BAYESIAN GLS REGRESSION FOR REGIONALIZATION OF HYDROLOGIC STATISTICS, FLOODS AND BULLETIN 17 SKEW

A Thesis

Presented to the Faculty of the Graduate School

of Cornell University

In Partial Fulfillment of the Requirements for the Degree of

Master of Science

by

Andrea Gruber Veilleux

August 2009

ABSTRACT

The research presented in this thesis develops new statistical techniques for estimating regional skewness coefficients to improve flood frequency analysis in the United States. Flood frequency guidelines for the United States, specified in *Bulletin 17B*, recommend fitting the log-Pearson Type III (LP3) distribution to the series of annual flood maxima, in which the third moment of the distribution, the skewness coefficient γ, is combined with a regional skewness coefficient to improve its precision. The research presented here extends the quasi-analytic Bayesian analysis of the Generalized Least Squares (GLS) regional hydrologic regression framework introduced by Reis *et al*. [2005] to more accurately and precisely estimate regional skewness coefficients. Specifically, formulas derived within a Bayesian regression framework for the computation of estimators, standard errors, and diagnostic statistics are provided by Reis [2005] and Reis *et al.* [2005]. Diagnostic statistics further developed here include a Bayesian plausibility value, pseudo adjusted R-squared, pseudo-Analysis of Variance table, two diagnostic error variance ratios, as well as leverage and influence metrics. In addition, this research also develops a new σ-influence diagnostic statistic which, in conjunction with the Bayesian extension of GLS leverage and influence metrics, can be used to better identify rogue observations and to effectively address lack-of-fit when estimating skewness coefficients.

Currently, *Bulletin 17B* allows for regional skew values to be obtained from the skew map included with the Bulletin. As it is over 30 years old, the regional skew values from the *Bulletin 17B* skew map do not reflect annual maximum data acquired since 1976. This increase in available data, along with advances in computing power to support the Bayesian GLS regional hydrologic regression framework, allow for a

much more precise estimate of the regional skewness coefficient for use in flood frequency analysis.

This research employs the Bayesian GLS regression framework to estimate regional log-space skewness coefficients for three data sets: the Illinois River basin, the state of South Carolina, and the Southeastern United States. *Bulletin 17B* allows for the generation of skew prediction equations as an alternative method for determining regional skew coefficients when the mean squared error of the equations is smaller than reported from the Bulletin's skew map. These skew prediction equations can be generated using Ordinary Least Squares analysis, Weighted Least Squares analysis, Generalized Least Squares analysis employing the method of moment model-error-variance estimator introduced by Stedinger and Tasker [1985, 1986ab], or the new Bayesian GLS estimator. The advantages of using the Bayesian GLS estimation technique to determine a skew prediction equation are demonstrated here in the Illinois River basin and the state of South Carolina studies.

To correctly analyze the Southeastern United States data set, methods are developed for identifying and screening redundant sites corresponding to nested watersheds with similar drainage areas. Special attention is devoted to developing an improved cross-correlation model of annual peak flows. The Bayesian GLS analysis using 342 stations from the Southeastern U.S. results in a highly accurate, constant regional skew model ($\hat{\gamma} = -0.019$), with an average variance of prediction equal to 0.14. More complex models which include regional information and basin characteristics as additional regression parameters result in very little improvement. The application of the Bayesian estimator in the Southeastern study generates improved results over the mean square error of 0.30 reported for the *Bulletin 17B* regional map skew.

BIOGRAPHICAL SKETCH

Andrea Marie Gruber was born June 8, 1984 in Fountain Valley, California. She grew up in sunny Southern California with her parents, brother, and sister. Andrea graduated cum laude from Cornell University in May 2006 with a B.S. in Civil and Environmental Engineering. Up to the challenge of more Ithaca winters, Andrea immediately began pursuing her M.S. degree at Cornell in the department of Civil and Environmental Engineering with a concentration in Environmental and Water Resources Systems. On October 18, 2008, she married fellow Cornell engineering graduate student Michael Veilleux and took the name Andrea Gruber Veilleux.

To my parents, I wouldn't be here without you

ACKNOWLEDGMENTS

I owe thanks to many people who have provided encouragement and assistance as I worked to complete this thesis. First and foremost, I would like to thank my research advisor and special committee chair, Professor Jery R. Stedinger. I so appreciate not only his expert guidance and support, but also the endless enthusiasm he has for his work. His valuable wisdom has been instrumental in fostering an incredibly rewarding learning experience. I would also like to thank my minor advisor and special committee member, Professor Wilfried Brutsaert for his counseling and insight.

I sincerely acknowledge the support provided by a Water Resources Institute Internship Award #07HQAGOI61 by the U.S. Geological Survey, U.S. Department of the Interior. I am grateful to Dr. Timothy Cohn for the knowledgeable input he provided regarding the research presented in this thesis. I would also like to thank my friend and colleague Dr. Ken Eng for assisting me in my research and for continually encouraging me to think big.

Most importantly, I would like to thank my family for always being there for me. I wouldn't be who I am today without the love and support of my parents Evan and Jane, my brother Aaron, and my sister Allison. And most of all to my best friend and husband Mike, thank you for helping me through the stressful days with your gourmet meals.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

The research presented in this thesis develops new statistical techniques for estimating regional skewness coefficients to improve flood frequency analysis in the United States. Flood frequency guidelines for the United States, specified in *Bulletin 17B*, recommend fitting the log-Pearson Type III (LP3) distribution to the series of annual flood maxima, in which the third moment of the distribution, the skewness coefficient, is combined with a regional skewness coefficient to improve its precision. The research presented here extends the quasi-analytic Bayesian analysis of the Generalized Least Squares (GLS) regional hydrologic regression framework introduced by Reis *et al*. [2005] to more accurately and precisely estimate regional skewness coefficients. Specifically, formulas derived within a Bayesian regression framework for the computation of estimators, standard errors, and diagnostic statistics are provided by Reis[2005] and Reis *et al.* [2005]. Diagnostic statistics further developed here include a Bayesian plausibility value, pseudo adjusted R-squared, pseudo-Analysis of Variance table, two diagnostic error variance ratios, as well as leverage and influence metrics. In addition, this research also develops a new σ-influence diagnostic statistic which, in conjunction with the Bayesian extension of GLS leverage and influence metrics, can be used to better identify rogue observations and to effectively address lack-of-fit when estimating skewness coefficients. This framework is applied to three different data sets from different parts of the United States: the Illinois River Basin data set, the state of South Carolina data set, and the Southeastern United States data set, to develop regional skewness estimators for use in flood frequency analysis.

Chapter 1 of this thesis begins with Section 1.1 which provides a background on United States flood flow frequency procedures, highlights the importance of the skewness coefficient in determining estimates of flood quantiles, and clarifies the need for an improved regional skewness estimator. Next, Section 1.2 presents justification for a Bayesian Generalized Least Squares framework for regional hydrologic regression analyses. Finally, Section 1.3 offers a detailed outline of the body of the thesis.

1.1 United States Flood Flow Frequency Procedures and Regional Skew

In 1967 the United States Water Resources Council published *Bulletin 15* entitled "A Uniform Technique for Determining Flood Flow Frequencies" in response to an increased social interest in managing flood loss and decreasing flood risk within the United States. *Bulletin 15* addressed the need to coordinate flood frequency methods used by different federal agencies and it was expected that this bulletin would be used as a guideline for all U.S. federal agencies performing flood frequency analysis in the United States. *Bulletin 15* was later updated in 1976 (*Bulletin 17*), with a minor revision in 1977 (*Bulletin 17A*), and a larger revision in 1982 to yield the current version, *Bulletin 17B*. An update of the U.S. recommended flood frequency guidelines has not occurred since 1982, thus almost 26 years has passed without revisions. As argued by Stedinger and Griffis [2008], it is essential that the prescribed, techniques in *Bulletin 17B* be updated in order to make use of recent advances in the field of flood frequency analysis.

The recommended approach for flood frequency analysis, as presented in *Bulletin 17B*, is to fit a log-Pearson Type III (LP3) distribution to the series of annual maximums. This distribution, in the specific case of flood frequency analysis, is described by three moments: the mean, the standard deviation, and the skewness

coefficient of the logarithms of the flow. The third moment, the skewness coefficient, is a measure of the asymmetry of the distribution or, in other words, the relative thickness of the tails of the distribution. The skewness coefficient is very sensitive to extreme events, such as large floods, as they cause a sample to be highly skewed, or asymmetrical. Thus, in flood frequency analysis, the skewness coefficient becomes significant because interest is focused on the right hand tail of the distribution. However, the span of available years of recorded flood data at a given gauge site is usually too short to provide a highly reliable estimate of the skewness coefficient. In order to improve the precision of the skewness estimator, *Bulletin 17B* advises combining a regional skew with the at-site skew [Beard 1974; Griffis and Stedinger, 2007b ; Hardison, 1975; IACWD, 1982; McCuen, 1979, 2001; Tasker, 1978]. Griffis and Stedinger [2009b, Appendix] show that the *Bulletin 17B* mean squared error (MSE) weighted skewness estimator results in the estimator with the smallest MSE provided that the regional skew is unbiased and independent of the at-site skew estimator. Griffis and Stedinger [2007a, 2009a] illustrate the value of a good regional skewness estimator in terms of the precision of flood quantile estimates.

When putting *Bulletin 17B* into practice, regional skew values may be obtained from the included skew map. The skew map in *Bulletin 17B*, Figure 1.1, is a slightly revised version of the map developed by Hardison [1975]. The skew map in *Bulletin 17B* is the same as the skew map generated in 1976 for *Bulletin 17* and is based on 2,972 stream gauging sites with records at least 25 years in length.

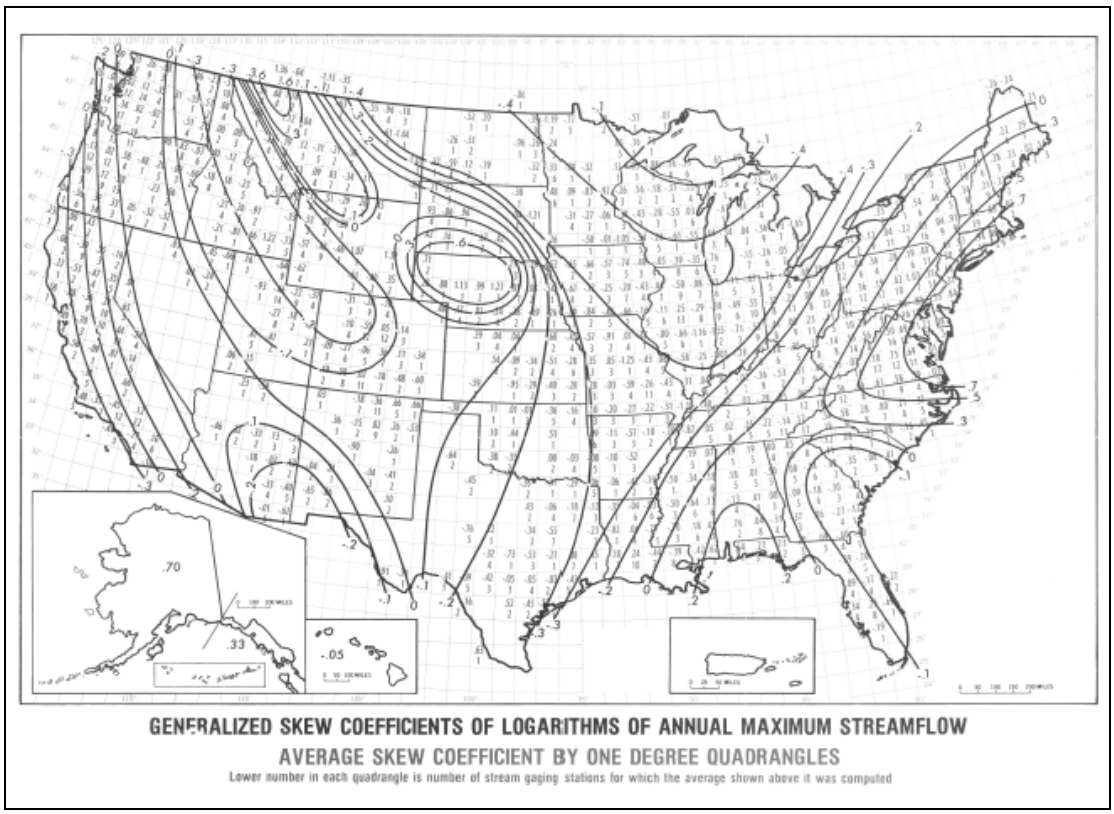

Figure 1.1: Plate I from *Bulletin 17B*: Map of Generalized Skew Coefficients of Logarithms of Annual Maximum Streamflow (IACWD, [1982])

This skew map is still used today, over 30 years later. The first edition of *Bulletin 17* states: "It is expected that Plate I [the skew map] will be revised as more data become available and more extensive studies are completed." However, in spite of the tremendous advances in computing power over the past few decades which support the Bayesian GLS regional hydrologic regression framework, the skew map has not recently been updated nationally and a revision of *Bulletin 17B* has not been published.

1.2 Regional Hydrologic Regression Analysis

Hydrologic studies that require information at ungauged sites pose a special challenge because no at-site information is available. Thus, there is a desire to develop regional hydrologic relationships based upon records available across a region. Also, at gauged sites, records can be too short to provide highly accurate at-site estimates of flood quantiles, low flows and other regional hydrologic statistics, and thus, regional information can also be of use to improve accuracy of estimates in these cases. One approach for relating data from gauged sites to ungauged sites is to derive an empirical relationship between the hydrologic variable of interest and various measurable basin characteristics at the gauged sites using regional regression analysis.

For many years, regional regression analysis used an Ordinary Least Squares (OLS) framework that considers the residual errors to be homoscedastic and independently distributed [Riggs, 1973]. However, the estimates of the variable of interest at different gauged sites have different precision due to differences in record length [Tasker, 1980; Kuczera, 1983] and possible differences in the precision of measurements and their variability [Tasker and Stedinger, 1989]. Stedinger and Tasker [1985,1986 ab] developed a GLS framework, which considers both differences in record lengths and precisions, as well as cross-correlation among station estimators since station estimators are generally correlated. This spatial correlation arises due to the fact that basins in close proximity to one another can experience their maximum flows from the same hydrologic event, so that the records upon which flow statistics are computed are correlated, resulting in cross-correlated streamflow statistics. Stedinger and Tasker showed that a GLS analysis provides better estimates of the model parameters and the model error variance in terms of mean squared errors than does an OLS approach. (See also Kroll and Stedinger, 1998)

GLS regression has been widely used, in both the United States and internationally, in many hydrologic studies, including the regionalization of flood quantiles, water quality parameters, low-flow statistics, and extreme rainfall [Tasker *et al*., 1986; Curtis, 1987; Tasker and Driver, 1988; Landers and Wilson, 1991; Moss and Tasker, 1991; Ludwig and Tasker, 1993; Rosbjerg and Madsen, 1995; GREHYS, 1996; Madsen and Rosbjerg, 1997; Robson and Reed, 1999; Kjeldsen and Rosbjerg, 2002; Feaster and Tasker, 2002; Madsen *et al*., 2002; Miceuski and Kuczera, 2009]. A Weighted Least Squares (WLS) procedure, which considers only differences in record lengths, has been used for the regionalization of the shape parameter (the skewness coefficient) by Tasker and Stedinger [1986], and for the state of Kansas [Rasmussen and Perry, 2000] and the state of North Carolina [Pope *et al*., 2001]. Various studies using region-of-influence techniques to regionalize flood quantiles have used GLS as a regression method [Tasker *et al*., 1996; Law and Tasker, 2003, Eng *et al*., 2007a, Eng *et al*. 2007b]. Moreover, a GLS analysis has also been used as the basis of hydrologic network design [Tasker, 1986; Medina, 1987; Tasker and Stedinger, 1989; Moss and Tasker, 1991; Soenksen *et al*., 1999].

Reis *et al*. [2003, 2005] introduced a Bayesian approach to parameter estimation for the Generalized Least Squares (GLS) regression analysis developed by Stedinger and Tasker [1985, 1986ab] for regional hydrologic analysis. A Bayesian analysis [Zellner, 1971; Gelman *et al*., 2004] provides both an exact measure of precision of the model error variance that method of moment (MM) and maximum likelihood (ML) estimators lack, and a more reasonable description of the possible values of the model error variance in cases where the MM and ML model error variance estimators are zero or nearly zero [Madsen and Rosbjerg, 1997]. The results presented in Reis *et al*. [2005] show that for cases in which the model error variance is small compared to the sampling error of the at-site estimates, which is often the case

for regionalization of the skewness coefficient, the Bayesian posterior distribution provides a more reasonable description of the model error variance than both the MM and ML point estimators. The MM estimator of the model error variance can be zero if the observed variability in the data is explained by the sampling error in the at-site estimates, causing a distortion in the uncertainty of the regional estimate. Similarly, the ML estimator of the model error variance may not be a good representation of the possible values of the model error variance when its value is small or zero because the likelihood function is often highly skewed; this results in the mode being a less appropriate summary statistic than the center-of-mass. Sometimes, the mode is at the origin which results in a ML estimate of zero even through routine values are very likely.

Qian *et al*. [2005] employ a similar Bayesian analysis for a watershed-loading model with three error terms representing independent observational errors, a structural correlated spatial dependency, and the impact of errors in one reach on the distribution of the estimated loads downstream. Jeong *et al*. [2007] used Bayesian GLS analysis for regionalization of the coefficient of L-moment variation (L-CV) and L-moment skew (L-Skew) in Korea for the Generalized Extreme Value (GEV) distribution.

Reis *et al.* [2005] and Reis [2005] developed a Bayesian GLS (B-GLS) framework together with diagnostic statistics for use in the estimation of regional skewness coefficients. This framework is discussed in detail in Chapter 2 as it pertains to the regionalization of hydrologic data. The diagnostic statistics developed and presented by Reis *et al*. [2005], Reis [2005], and Griffis and Stedinger [2007b] include: the average variance of prediction for a new site (AVP_{new}), Bayesian plausibility, error variance ratio (EVR), misrepresentation of the beta variance (MBV), R_δ^2 and pseudo analysis of variation (pseudo ANOVA), leverage, and influence. The

AVP_{new} and the Bayesian plausibility value can guide model selection. EVR calculates whether a WLS or GLS analysis is likely to be needed, or if an OLS analysis will suffice. Similarly, MBV indicates if GLS analysis is needed, or WLS analysis will suffice. The R_{δ}^2 statistic describes how well the model is explaining the variability in the true dependent variable, while the pseudo ANOVA table describes how much of the variation in the observations can be attributed to the model, and how much to the model error and sampling error. Finally, leverage and influence metrics identify and consider the impact of unusual observations on the models. Those metrics, including the newly developed σ-influence introduced by Gruber *et al*. [2007], allow for a comprehensive examination of a regression analysis developed within the B-GLS framework.

1.3 Thesis Organization

Chapter 2 of this thesis develops the Bayesian Generalized Least Squares framework and diagnostic statistics for regional hydrologic regression analysis. Chapter 3 then compares the results of OLS, WLS, and GLS analyses, as well as evaluates the use of method of moment versus Bayesian estimators to derive regional models of the skewness coefficient of the log-Pearson Type III distribution for two data sets: the Illinois River basin (62 sites) and the state of South Carolina (89 sites). The earlier paper by Reis *et al*. [2005] similarly analyzed data for the Tibagi River basin (17 sites) and the Muskingum River basin (44 sites).

Chapter 4 discusses a study employing B-GLS regionalization techniques and diagnostic statistics within a (489 site) data set representing seven Southeastern U.S. states [Gruber and Stedinger, 2008]. An important part of this study is the selection of annual peak flow data for use in the B-GLS regression framework and the estimation of cross-correlations of annual peaks. As this large study indicates that many gauge

site records are from watersheds largely contained within another watershed represented by a different gauge site, it considers the impact of such nested watersheds on regional studies, and develops criteria for identifying redundant gauge sites. Further, different models for the cross-correlations of annual peak flows between two sites, an integral part of the B-GLS regression framework, are explored. After considering the aforementioned data-related issues resulting from such a large Southeastern U.S. study area, a regional skew estimator is developed using B-GLS regression. The case study illustrates the use and value of the different leverage and influence statistics.

Finally, Chapter 5 describes the accomplishments of this research focusing on regional skewness coefficients. In particular the quasi-analytic Bayesian analysis of a GLS regression model described by Reis *et al.* [2005] is now an operational GLS regional hydrologic regression methodology. The research documented in this thesis provides examples that illustrate both the performance of the Bayesian GLS analysis in the estimation of regional skewness coefficients and the value of the diagnostic statistics.

REFERENCES

Beard, L. R., (1974), *Flood Flow Frequency Techniques*, Center for Research in Water Resources, The University of Texas at Austin.

Curtis, G.W., (1987), Technique for estimating flood-peak discharges and frequencies on rural streams in Illinois, U.S. Geological Survey Water-Resources Investigations Report 87-4207.

Eng, K., Milly, P.C.D., and Tasker, G.D., (2007a), Flood regionalization: A hybrid geographic and predictor-variable region-of-influence regression method: Journal of Hydrologic Engineering, v. 12, p. 585 - 591.

Eng, K., Stedinger, J.R., and Gruber, A.M., (2007b), Regionalization of streamflow characteristics for the Gulf-Atlantic Rolling Plains using leverage guided region-of -influence regression, in Kabbes, K.C., ed., Proceedings of the World Environmental and Water Resources Congress, May 15–19, 2007, Tampa, Florida, USA: American Society of Civil Engineers.

Feaster, T.D. and G.D. Tasker, (2002), Techniques for Estimating the Magnitude and Frequency of Floods in Rural Basins of South Carolina, 1999, Water Resources Investigations Report 02-4140, U.S. Geological Survey: Columbia, South Carolina.

Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B., (2004), *Bayesian Data Analysis*, Chapman & Hall/CRC, Boca Raton, FL.

GREHYS, (1996), Presentation and review of some methods for regional flood frequency analysis. *J. Hydrol.* 186 1–4, pp. 63–84.

Griffis, V.W., and J. R. Stedinger, (2007a), The LP3 distribution and its application in flood frequency analysis, 2. Parameter Estimation Methods, *J. of Hydrol. Engineering,* 12(5), 492-500.

Griffis, V. W., and J. R. Stedinger, (2007b), The Use of GLS Regression in Regional Hydrologic Analyses, J. of Hydrology, 344(1-2), 82-95, [doi:10.1016/j.jhydrol.2007.06.023].

Griffis, V.W., and J. R. Stedinger, (2009a), Closure: The LP3 distribution and its application in flood frequency analysis, 2. Parameter Estimation Methods, *J. of Hydrol. Engineering 14(2)*, 209-212.

Griffis, V.W., and J. R. Stedinger, (2009b), Log-Pearson Type 3 Distribution and Its Application in Flood Frequency Analysis. III: Sample Skew and Weighted Skew Estimators, *J. of Hydrol. Engineering 14(2)*, 121-120, [doi:10.1061/(ASCE)1084-0699(2009)14:2(121)]

Gruber, Andrea M., Dirceu S. Reis Jr., and Jery R. Stedinger, (2007), Models of Regional Skew Based on Bayesian GLS Regression, Paper 40927-3285, World Environmental & Water Resources Conference - Restoring our Natural Habitat, K.C. Kabbes editor, Tampa, Florida, May 15-18.

Gruber, Andrea M. and Jery R. Stedinger, (2008), Models of LP3 Regional Skew, Data Selection and Bayesian GLS Regression, Paper 596, World Environmental and Water Resources Congress – Ahupua'a, Babcock, R.W. and R. Watson editors, Honolulu, Hawai'I, May 12-16.

Models of Regional Skew Based on Bayesian GLS Regression, Paper 40927-3285, World Environmental & Water Resources Conference - Restoring our Natural Habitat, K.C. Kabbes editor, Tampa, Florida, May 15-18.

Hardison, C.H., (1975), Generalized skew coefficients of annual floods in the United States and their application, *Water Resour Res.* 11(6), 851-854.

Interagency Committee on Water Data (IACWD). (1982). *Guidelines for determining flood flow frequency: Bulletin 17B (revised and corrected)*, Hydrol. Subcomm., Washington, D.C., 28.

Jeong, Dae Il, Jery R. Stedinger, Young-Oh Kim, and Jang Hyun Sung, (2007), Bayesian GLS for Regionalization of Flood Characteristics in Korea, Paper 40927-2736, World Environmental & Water Resources Conference - Restoring our Natural Habitat, K.C. Kabbes editor, Tampa, Florida, May 15-18.

Kjeldsen, T.R. and D. Rosbjerg, (2002), Comparison of regional index flood estimation procedures based on the extreme value type I distribution, *Stoch. Env. Res. Risk A.,* 16(5), 358-373.

Kroll, C.N., and J.R. Stedinger, (1998), Regional hydrologic analysis: Ordinary and generalized least squares revisited, *Water Resour. Res. 34*(1), 121-128.

Kuczera, G., (1983), "Effect of Sampling Uncertainty and Spatial Correlation on an Empirical Bayes Procedure for Combining Site and Regional Information," *Journal of Hydrology*, 65, 373-398.

Landers, M.N. and K.V. Wilson, Jr., (1991), Flood Characteristics of Mississippi Streams, Water Resources Investigations Report 91-4037, U.S. Geological Survey in cooperation with Mississippi State Highway Department, Jackson, Mississippi.

Law, G.S., and G.D. Tasker, (2003), Flood-frequency prediction methods for unregulated streams of Tennessee, 2000, WRIR 03-4176.

Ludwing, A.H. and G.D. Tasker, (1993), "Regionalization of Low-Flow Characteristics of Arkansas Streams," U.S. Geological Survey Water-Resources Investigations Report 93-4013.

Madsen, H., and D. Rosbjerg, (1997), Generalized least squares and empirical Bayes estimation in regional partial duration series index-flood modeling, *Water Resour Res., 33*(4), 771-782.

Madsen, H., P. S. Mikkelsen, D. Rosbjerg, and P. Harremoe¨s, (2002), Regional estimation of rainfall intensity-duration-frequency curves using generalized least squares regression of partial duration series statistics, Water Resour. Res., 38(11), 1239, doi:10.1029/2001WR001125.

McCuen, R.H., (1979), Map skew ???, *J. Wate Resour. Plan. And Manage. Div.*, ASCE, 105(WR2), 265-277 [with Closure 107(WR2), 582, 1981].

McCuen, R.H., (2001), Generalized flood skew: map versus watershed skew, *J. Hydrologic Eng.*, ASCE, *Vol*. 6(4), 293-299.

Medina, K.D.. (1987), Analysis of Surface-Water Data Network in Kansas for effectiveness in providing regional streamflow information (with a section on "Theory and application of generalized least squares" by G.D. Tasker), USGS Water-Supply Paper 2303. Reston, VA: US Geological Survey. 28 pp.

Micevski, T., and G. Kuczera (2009), Combining site and regional flood information using a Bayesian Monte Carlo approach, *Water Resour. Res.*, 45, W04405, doi:10.1029/2008WR007173

Moss, M.E. and G.D. Tasker, (1991), "An intercomparison of hydrological network-design technologies," *Hydrological Sciences Journal*, 36(3), 209.

Pope, F., G.D. Tasker, J.C. Robbins, (2001), Estimating the Magnitude and Frequency of Floods in Rural Basins of North Carolina – Revised, Water Resources Investigations Report 01-4207, U.S. Geological Survey, Raleigh, North Carolina.

Qian, S. S.; Reckhow, K. H.; Zhai, J., McMahon, G., (2005), Nonlinear regression modeling of nutrient loads in streams: A Bayesian approach, *Water Resour. Res., 41*(7), W07012*,* doi: 10.1029/2005WR003986, 2005.

Rasmussen, P.P. and C.A. Perry, (2000), Estimation of Peak Streamflows for Unregulated Rural Streams in Kansas, Water Resources Investigations Report 00-4079, U.S. Geological Survey, Columbia, Lawrence, Kansas.

Reis Jr., D.S., (2005). Flood Frequency Analysis Employing Bayesian Regional Regression and Imperfect Historical Information. Ph.D. Dissertation, Cornell University.

Reis, D. S., Jr., J.R. Stedinger, and E.S. Martins, (2003), Bayesian GLS Regression with application to LP3 Regional Skew Estimation, Proceedings World Water & Environmental Resources Congress 2003, Editors P. Bizier and P. DeBarry, Philadelphia, PA, American Society of Civil Engineers, June 23-26.

Reis, D. S., Jr., J. R. Stedinger, and E. S. Martins, (2005), Bayesian generalized least squares regression with application to log Pearson type 3 regional skew estimation, *Water Resour. Res.*, *41*, W10419, doi:10.1029/2004WR003445.

Riggs, H. C., (1973), Regional Analyses of Streamflow Characteristics: Techniques of Water-Resources Investigations of the United States Geological Survey, Book 4, Chapter B3.

Robson, Alice and Duncan Reed, (1999), Flood Estimation Handbook Volume 3: Statistical procedures for flood frequency estimation, Institute of Hydrology, Oxfordshire, UK.

Rosbjerg, D. and H. Madsen, (1995), Uncertainty measures of regional flood frequency estimators, *Journal of Hydrology*, 167, 209-224.

Soenksen, Philip J., L. D. Miller, J. B. Sharpe, J. R. Watton, (1999), Peak-flow frequency relations and evaluation of the peak-flow gaging network in Nebraska, Water-Resources Investigation Report 99-4032, U.S. Geological Survey, Lincoln, Nebraska.

Stedinger, J.R. and V.W. Griffis, (2008), Flood Frequency Analysis in the United States: Time to Update. (editorial) *J. of Hydrol. Engineering*, April, pp. 199-204.

Stedinger, J.R., and G.D. Tasker, (1985), Regional Hydrologic Analysis, 1. Ordinary, Weighted and Generalized Least Squares Compared, *Water Resources Research, 21*(9), 1421-1432.

Stedinger, J.R. and G. Tasker, (1986a), Correction to "Regional hydrologic analysis, 1, Ordinary, weighted and generalized least squares compared", *Water Res. Research*, 22(5), 844.

Stedinger, J.R. and G. Tasker, (1986b), Regional hydrologic analysis, 2: Model-error estimators, estimation of sigma and log-Pearson Type 3 distributions, *Water Res. Research*, 22(10), 1487-1499.

Tasker, G.D., (1978), Flood frequency analysis with a generalized skew coefficient, *Water Resources Research*, 14(2), 373-376.

Tasker, G.D., (1980), "Hydrologic Regression with Weighted Least Squares," *Water Resources Research, 16*(6), 11107-1113.

Tasker, G.D., (1986), "Generating efficient gaging plans for regional information," Integrated Design of Hydrologic Networks, IAHS Publ. no. 158.

Tasker, G.D., and J.R. Stedinger, (1986), Estimating Generalized Skew With Weighted Least Squares Regression, Journal of Water Resources Planning and Management, 112(2), 225-237.

Tasker, G.D., and J.R. Stedinger, (1989), An Operational GLS Model for Hydrologic Regression, *Journal of Hydrology, 111*(1-4), 361-375.

Tasker, G.D. and N.E. Driver, (1988), "Nationwide Regression Model for Predicting Urban Runoff Water Quality at Unmonitored Sites," Water Resources Bulletin 24(5), 1091-1101.

Tasker, G.D., J.H. Eychaner and J.R. Stedinger, (1986), Application of Generalized Least Squares in Regional Hydrologic Regression Analysis, in Selected Papers in the Hydrological Science, U.S. Geological Survey, Water Resources Division, Reston, VA, Water Supply Paper 2310, pp. 107-116, December.

Tasker, G.D., S.A. Hodge, and C.S. Barks, (1996), "Region of Influence Regression for Estimating the 50-Year Flood at Ungauged Sites," Water Resources Bulletin 32(1), 163-170.

Zellner, A., (1971), *An Introduction to Bayesian Inference in Econometrics*, John Wiley and Sons, Inc., New York.

CHAPTER 2

DEVELOPMENT OF MODELS OF REGIONAL SKEW BASED ON BAYESIAN GENERALIZED LEAST SQUARES FRAMEWORK

2.1 Introduction to Bayesian- Generalized Least Squares Framework

2.1.1 Generalized Least Squares (GLS) Regression for Hydrologic Statistics

Streamflow data sets can be used to derive an empirical relationship between hydrologic characteristics at a site, such as the T-year flood or log-space skewness coefficient used to fit a log-Pearson Type III distribution, and physiographic variables, such as drainage area and channel slope. This GLS analysis, as described by Stedinger and Tasker [1985, 1986a] and Tasker and Stedinger [1989], assumes that the actual value of the quantity of interest y_i (or some transformation) for a given site i can be described by a function of physiographic characteristics with an additive error

$$y_i = \beta_0 + \sum_{j=1}^{k} \beta_j X_{ij} + \delta_i \qquad i = 1, 2, \ldots, n \text{ stations} \tag{2.1}$$

wherein X_{ij} (j=1…k) are the elements of a matrix of k explanatory variables based upon the physical characteristics at each site i, β_i are the model parameters, and δ_i are the independently distributed model errors with the following properties:

$$\mathrm{E}\left[\delta_i\right] = 0\,, \qquad \mathrm{Cov}\left(\delta_i, \delta_j\right) = \begin{cases} \sigma_\delta^2 & i = j \\ 0 & i \neq j \end{cases} \tag{2.2}$$

However, in most analyses, only an at-site estimate of $\hat{y}_i$ is available and thus a time sampling error η_i, should be introduced into the model, such that

$$\hat{y}_i = y_i + \eta_i \qquad i = 1, 2, \ldots, n \text{ stations} \tag{2.3}$$

with

$$\mathrm{E}[\eta_i] = 0 \qquad \mathrm{Cov}(\eta_i, \eta_j) = \begin{cases} \sigma^2_{\eta_i} & i = j \\ \sigma_{\eta_i}\sigma_{\eta_j}\rho_{ij} & i \neq j \end{cases} \tag{2.4}$$

wherein $\sigma^2_{\eta_i}$ is the at-site sampling error variance for $\hat{y}_i$, and ρ_{ij} is the sampling error correlation coefficient due to correlation among the statistic of interest at stations *i* and *j* (cross-site correlation).

In matrix notation, the GLS model is

$$\hat{\mathbf{y}} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\eta} + \boldsymbol{\delta} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{2.5}$$

where the (n x k+1) matrix $\mathbf{X}$ contains ones in the first column and values of the k explanatory variables in the remaining columns, the vector $\boldsymbol{\beta}$ has the (k+1) parameters of the model that must be estimated, the vector $\boldsymbol{\eta}$ contains the sampling errors in the sample estimators, and the vector $\boldsymbol{\delta}$ contains the model errors for the n sites used in the analysis.

The errors ε_i are a combination of: (i) time-sampling-error η_i in the sample estimators of y_i and (ii) underlying model error δ_i. The total error vector $\boldsymbol{\varepsilon}$ has mean zero and covariance matrix

$$E\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T\right] = \boldsymbol{\Lambda}(\sigma^2_\delta) = \sigma^2_\delta \mathbf{I} + \boldsymbol{\Sigma}(\hat{\mathbf{y}}) \tag{2.6}$$

where $\boldsymbol{\Sigma}(\hat{\mathbf{y}})$ is the covariance matrix of the sampling errors in the sample estimators whose elements are given by equation (2.4), and σ^2_δ is the underlying model error variance, which must be determined. The value of σ^2_δ can be viewed as a heterogeneity measure [Madsen and Rosbjerg, 1997; Madsen et al., 2002].

Weighted Least Squares (WLS) and Ordinary Least Squares (OLS) analyses are special cases of a GLS analysis. When $\hat{\rho}\left(\hat{y}_i, \hat{y}_j\right) = 0$ for every pair of sites ($i \neq j$), GLS reduces to WLS. WLS reduces to OLS when the diagonal covariance matrix has elements on the diagonal equal to a common value.

The GLS estimator of $\boldsymbol{\beta}$ and its respective covariance matrix for known σ_δ^2 are given by

$$\mathbf{b} = \left[\mathbf{X}^T \boldsymbol{\Lambda}(\sigma_\delta^2)^{-1}\mathbf{X}\right]^{-1} \mathbf{X}^T \boldsymbol{\Lambda}(\sigma_\delta^2)^{-1}\hat{\mathbf{y}} \tag{2.7a}$$

$$\boldsymbol{\Sigma}[\mathbf{b}] = \left[\mathbf{X}^T \boldsymbol{\Lambda}(\sigma_\delta^2)^{-1}\mathbf{X}\right]^{-1} \tag{2.7b}$$

The model error variance σ_δ^2 can be estimated by either generalized method of moment (MM) or maximum likelihood (ML) estimators, as described by Stedinger and Tasker [1986b]. The MM generalized estimator is determined by iteratively solving equation (2.7a) along with the generalized residual mean square error equation:

$$(\hat{\mathbf{y}} - \mathbf{X}\mathbf{b})^T \left[\hat{\sigma}_\delta^2 \mathbf{I} + \boldsymbol{\Sigma}(\hat{\mathbf{y}})\right]^{-1} (\hat{\mathbf{y}} - \mathbf{X}\mathbf{b}) = n - (k+1) \tag{2.8}$$

for n sites and k+1 parameters. In some situations, the sampling covariance matrix explains all the variability observed in the data, which means the left-hand side of equation (2.8) will be less than $n - (k+1)$ even if $\hat{\sigma}_\delta^2$ is zero. In these circumstances, the MM estimator of the model error variance is generally taken to be zero [Stedinger and Tasker, 1985, 1986b].

The ML estimators of $\boldsymbol{\beta}$ and σ_δ^2 can be obtained by minimizing the negative of the log-likelihood function of the residuals, which are assumed to be normally distributed with zero mean and the covariance matrix in equation (2.6):

$$\min\left\{\ln\left[\det\left(\boldsymbol{\Lambda}\left(\sigma_\delta^2\right)\right)\right] + (\hat{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta})^T \boldsymbol{\Lambda}\left(\sigma_\delta^2\right)^{-1} (\hat{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta})\right\} \tag{2.9}$$

subject to $\hat{\sigma}_\delta^2 \geq 0$. The ML estimate of $\boldsymbol{\beta}$ is the same as the one computed by using equation (2.7), except that the value of $\hat{\sigma}_\delta^2$ will be different. The variance of $\boldsymbol{\beta}$ is the same as in equation (2.7) because $\boldsymbol{\beta}$ and $\hat{\sigma}_\delta^2$ are asymptotically independent [Rencher, 2000]. The inverse of the second derivative of the likelihood function could be used to estimate the variance of the model error variance estimator when the constraint $\hat{\sigma}_\delta^2 \geq 0$ is not binding [Bickel and Docksum, 1977].

2.1.2 Bayesian Approach

Reis *et al*. [2005] develop a Bayesian analysis of the GLS model. In particular, they compute the posterior moments of the $\boldsymbol{\beta}$ parameters and the full posterior distribution of the model error variance σ_δ^2.

The Bayesian approach requires the specification of prior distributions for both the $\boldsymbol{\beta}$ parameters and model error variance σ_δ^2. A multivariate normal distribution with a mean of zero and a large variance is used for the prior distribution for $\boldsymbol{\beta}$. This almost non-informative prior distribution produces a probability density function (pdf) that is relatively flat in the region of interest. The prior information for the model error variance σ_δ^2 is represented by an exponential distribution with parameter λ, which represents the reciprocal of the prior mean of the model error variance. Following Reis *et al*. [2005] for the regionalization of skews, a value of λ equal to 6 is employed, though as experience accumulates in the future a smaller value or a different distribution may be justified.

The likelihood function for the data $\hat{\mathbf{y}}$ is considered to be a multivariate normal distribution; thus, the marginal posterior distribution of the model error variance can be computed by integrating the joint posterior distribution over the possible values of $\boldsymbol{\beta}$ [Zellner, 1971, eqn. 8.14; Kitanidis, 1986] to obtain

$$f(\sigma_\delta^2 \mid \hat{\mathbf{y}}) = \int f(\boldsymbol{\beta}, \sigma_\delta^2 \mid \hat{\mathbf{y}}) d\boldsymbol{\beta} \propto \int f(\hat{\mathbf{y}} \mid \boldsymbol{\beta}, \sigma_\delta^2) \;\; \xi(\boldsymbol{\beta}, \sigma_\delta^2) d\boldsymbol{\beta} \tag{2.10}$$

where $f(\boldsymbol{\beta}, \sigma_\delta^2 \mid \hat{\mathbf{y}})$ is the joint posterior of the parameters, $f(\hat{\mathbf{y}} \mid \boldsymbol{\beta}, \sigma_\delta^2)$ is the likelihood function for the data $\hat{\mathbf{y}}$, and $\xi(\boldsymbol{\beta}, \sigma_\delta^2)$ is the joint prior for $\boldsymbol{\beta}$ and σ_δ^2. If one uses a non-informative prior on $\boldsymbol{\beta}$, the marginal posterior distribution for the model error variance, except for the normalizing constant, is

$$f(\sigma_\delta^2 \mid \hat{\mathbf{y}}) \propto \left[\left| \boldsymbol{\Lambda}(\sigma_\delta^2) \right| \left| \mathbf{X}^T \boldsymbol{\Lambda}(\sigma_\delta^2)^{-1} \mathbf{X} \right| \right]^{-1/2} \exp\left[-0.5 (\hat{\mathbf{y}} - \mathbf{X}\mathbf{b})^T \boldsymbol{\Lambda}(\sigma_\delta^2)^{-1} (\hat{\mathbf{y}} - \mathbf{X}\mathbf{b}) \right] \;\; \xi(\sigma_\delta^2) \tag{2.11}$$

wherein |**A**| denotes the determinant of a matrix **A**. With equation (2.11), one can compute the marginal pdf, mean, and variance of σ_δ^2 by a numerical evaluation of one-dimensional integrals [Reis *et al*., 2005]. Similarly, posterior moments of $\boldsymbol{\beta}$ can also be computed by a one-dimensional numerical integration using the pdf in equation (2.11) where the conditional distribution of $\boldsymbol{\beta}$ given σ_δ^2 is normal with mean and variance given in (2.7a,2.7b); thus yielding

$$\begin{aligned} \boldsymbol{\mu}_\beta &= E(\boldsymbol{\beta} \mid \hat{\mathbf{y}}) = \int \boldsymbol{\beta} f\left(\boldsymbol{\beta} \mid \hat{\mathbf{y}}\right) d\boldsymbol{\beta} \\ &= \int E(\boldsymbol{\beta} \mid \sigma_\delta^2, \hat{\mathbf{y}}) f(\sigma_\delta^2 \mid \hat{\mathbf{y}}) d\sigma_\delta^2 = \int \mathbf{b}\left(\sigma_\delta^2\right) f(\sigma_\delta^2 \mid \hat{\mathbf{y}}) d\sigma_\delta^2 \end{aligned} \tag{2.12}$$

$$\begin{aligned} Var(\boldsymbol{\beta} \mid \hat{\mathbf{y}}) &= \int\int \left(\boldsymbol{\beta} - \boldsymbol{\mu}_\beta\right)\left(\boldsymbol{\beta} - \boldsymbol{\mu}_\beta\right)^T f(\boldsymbol{\beta} \mid \sigma_\delta^2, \hat{\mathbf{y}}) f(\sigma_\delta^2 \mid \hat{\mathbf{y}}) d\boldsymbol{\beta} d\sigma_\delta^2 \\ &= \int \left\{ \left(\mathbf{b}\left(\sigma_\delta^2\right) - \boldsymbol{\mu}_\beta\right)\left(\mathbf{b}\left(\sigma_\delta^2\right) - \boldsymbol{\mu}_\beta\right)^T + \left(\mathbf{X}^T \boldsymbol{\Lambda}\left(\sigma_\delta^2\right)^{-1} \mathbf{X}\right)^{-1} \right\} f\left(\sigma_\delta^2 \mid \hat{\mathbf{y}}\right) d\sigma_\delta^2 \end{aligned} \tag{2.13}$$

Here the posterior variance of $\boldsymbol{\beta}$ equals the variance of the conditional mean $\mathbf{b}\left(\sigma_\delta^2\right)$ plus the average of the conditional variance of $\boldsymbol{\beta}$ for a given σ_δ^2 [Reis *et al*., 2005].

2.2 Model Selection Criteria and Regression Diagnostics

2.2.1 Introduction to Model Selection

A goal of model selection is to resolve which set of possible explanatory variables afford the most accurate prediction, while also searching for the simplest model possible. Several traditional statistics are available for model selection: coefficient of determination (R^2), likelihood ratios, Mallows C_p, Akaike Information Criterion (AIC), and the Bayesian Information Criterion (BIC) [Linhart and Zucchini, 1986; Gelman *et al*., 2004]. Qian *et al*. [2005] employ Bayes Factors and a Bayesian Deviance Information Criterion (DIC). Many of these statistics penalize model complexity, thus, a sufficient improvement in the model's prediction ability must result so as to support the inclusion of an additional independent variable. Below, B-GLS regression statistics are developed to guide model selection.

2.2.2 Average Variance of Prediction

Interest is focused on making predictions at gauged and ungauged sites, thus, a natural metric to evaluate a model is a variance of prediction which penalizes the inclusion of extra independent variables because it accounts for the sampling variance of the parameters [Carlin and Louis, 2000]. Since the variance of prediction generally depends upon the values of the independent variables at a given site, Tasker and Stedinger [1986] suggest the use of an Average Variance of Prediction (AVP), computed across the *x*-values of sites used in the regression. This implicitly assumes that these sites are representative of the sites at which predictions will be made.

For a new and perhaps ungauged site with a row vector of explanatory characteristics $\mathbf{x}_o$ and a *y*-value of y_o, the posterior expected value of y_o is $\mathbf{x}_o\boldsymbol{\mu}_\beta$, where $\boldsymbol{\mu}_\beta$ is the posterior expected value of $\boldsymbol{\beta}$. With a Bayesian analysis, the posterior sampling error variance for $\mathbf{x}_o\boldsymbol{\mu}_\beta$ is

$$Var(\mathbf{x}_o\boldsymbol{\mu}_\beta) = \underset{\sigma_\delta^2}{E}\left\{\mathbf{x_o}\left[(\mathbf{b}-\boldsymbol{\mu}_\beta)(\mathbf{b}-\boldsymbol{\mu}_\beta)^T + (\mathbf{X}^T\boldsymbol{\Lambda}\left(\sigma_\delta^2\right)^{-1}\mathbf{X})^{-1}\right]\mathbf{x_o}^T\right\} = \mathbf{x}_o Var[\boldsymbol{\beta}\,|\,\hat{\mathbf{y}}]\mathbf{x}_o^T \qquad (2.14)$$

The posterior variance of prediction for the unknown true value y_o associated with $\mathbf{x}_o$ is given by

$$VP = E\left[\left(y_0 - \mathbf{x}_0\boldsymbol{\mu}_\beta\right)^2\right] = E[\sigma_\delta^2] + Var(\mathbf{x}_o\boldsymbol{\mu}_\beta) = E[\sigma_\delta^2] + \mathbf{x}_o Var[\boldsymbol{\beta}\,|\,\hat{\mathbf{y}}]\mathbf{x}_o^T \qquad (2.15)$$

A measure of how well OLS, WLS and GLS regression analysis will predict a hydrologic statistic on average over a new region, whose $\mathbf{x}_o$ is like those in the $\mathbf{X}$ matrix, is the average variance of prediction for a new site AVP_{new}, introduced by Tasker and Stedinger [1986]. For a Bayesian analysis, as used in Reis *et al*. [2005],

$$\mathrm{AVP}_{new} = E[\sigma_\delta^2] + \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i Var[\boldsymbol{\beta}\,|\,\hat{\mathbf{y}}]\mathbf{x}_i^T \qquad (2.16)$$

However, if the prediction is for a site *i* that was used for estimation of the parameters of the model, the measure of prediction for such an old site, VP_{old}, appears in the appendix (A.13), requires an additional term. The Appendix provides derivations of the expressions used to calculate variance of predictions.

Generally, it is anticipated that models will be used at new sites, so AVP_{new} is the critical statistic for model selection. However, if a regional estimator is nearly as precise or is more accurate than the at-site estimator, then one may use the regional estimator instead, or a weighted average, in which case VP_{old} becomes relevant.

2.2.3 Bayesian Plausibility Value

It is common in classical statistics to perform a hypothesis test to check if a given parameter is statistically different from zero. If one cannot reject the null hypothesis that the value of the parameter is equal to zero, the variable whose

coefficient is being tested is generally dropped from the model. It is also common to report a P-value, which reflects the probability, under the null hypothesis, of computing a parameter value as large as or larger than the value obtained from the sample.

The Bayesian Plausibility Value, ψ, developed by Reis *et al.* [2005] and expanded on in Gruber *et al.* [2007] describes whether zero is a plausible value for each **β**-parameter in a regression analysis given the prior distribution and the data. As discussed by Lindley [1965] and Zellner [1971], given the Bayesian posterior pdf of **β** and the data available, one can construct a credible region for the regression parameters. A credible region summarizes the posterior belief about a parameter and can be the basis of a hypothesis test that concludes that a parameter is zero if zero is included in a 90% or a 95% credible region. This allows one to perform the equivalent of a classical hypothesis test within a Bayesian framework using the posterior distribution of each parameter, which also reflects the prior distribution.

Thus, the plausibility level for zero is defined as the smallest probability ψ such that zero is in a 100(1- ψ)% credible region for a parameter. The plausibility value is computed as

$$\psi = 2 \underset{\sigma_\delta^2}{E} \left\{ \Phi \left[-\upsilon \frac{b(\sigma_\delta^2)}{\sigma_b(\sigma_\delta^2)} \right] \right\} \tag{2.17}$$

wherein Φ is the standard normal cumulative distribution function (cdf), and the conditional mean $b(\sigma_\delta^2)$ and standard error $\sigma_b(\sigma_\delta^2)$ for β_i are both dependent on σ_δ^2 as indicated in equations (2.7a,2.7b); ν = sign[μ_β] = 1 for $\mu_\beta \geq 0$ and -1 for $\mu_\beta < 0$.

Bayarri and Berger [2000] and Robins *et al*. [2000] discuss the Bayesian P-value which corresponds to the probability that another random sample X would generate a more extreme value of a test statistic than that which is observed, and thus

is a statistic more consistent with the classical P-value. Those authors and others have tried to develop a Bayesian P-value that strictly reflects the data and not the prior distribution. In this research, the Bayesian Plausibility Value reflects the Bayesian point of view that the prior distribution is also information about the parameters, and thus, it is appropriate to use such information when deciding when to include a parameter in this model.

2.2.4 Introduction to Regression Diagnostics

The goal of regression diagnostics is to allow for a comprehensive examination of a regression analysis developed with the B-GLS framework. Upcoming sections propose a pseudo adjusted R^2, discuss analysis of variance for a GLS analysis, as well as present other diagnostic statistics including two error variance ratios, leverage, influence and σ-Influence.

2.2.5 R_δ^2 and Analysis of Variance

The traditional R^2 statistic measures the degree to which a model explains the variability in the data. It uses the partitioning of the sum of squared deviations and associated degrees of freedom to describe the variance of the signal versus the model error. Traditionally for OLS regression, the Total-Sum-of-Squared deviations about the mean (SST) is divided into two separate terms, the Sum-of-Squared Errors explained by the Regression model (SSR) and the residual Sum-of-Squared Errors (SSE), where SST = SSR + SSE [Devore, 2004]. The coefficient of determination R^2 and the adjusted R^2, (denoted $\bar{R}^2$), both describe the fraction of the total variability the model explains, computed as

$$R^2 = \frac{\text{SST} - \text{SSE}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}, \tag{2.18}$$

$$\bar{R}^2 = 1 - \frac{\mathrm{SSE}/(n-k-1)}{\mathrm{SST}/(n-1)} = 1 - \frac{s_\varepsilon^2}{s_y^2} \tag{2.19}$$

Here n is the total number of observations and k is the number of covariates used in the regression model in addition to a constant.

For WLS and GLS analyses, these formulas do not provide the intended insight. The error of most concern is the model error variance because the sampling error is unexplainable and represents noise that otherwise complicates the analysis. A new measure is needed in which the sampling error variance is separated from the total error variance, leaving behind the fraction of the variance accounted for by the model and by the model error. Griffis and Stedinger [2007] developed such a pseudo metric, R_{GLS}^2, to quantify the portion of the variance explained by the model.

Liu *et al*. [2005] and Han *et al*. [2009] use negative binomial regression to explain variations in the expected number of power distribution system failures. However, their observed discrete count data includes sampling variability, as does the $\hat{\mathbf{y}}$ in the B-GLS regional skew regression as shown in equation (2.3). Their pseudo-R^2 value describes the fraction of the true variability in the expected number of failures that their regression model explains.

The natural estimator for WLS and GLS models of the expected value of the sum of squared model errors would be $n\sigma_\delta^2(k)$ for a model with k covariates plus a constant. A corresponding estimator of the expected total variation in the true y_i values, corresponding to SST neglecting sampling error, is $n\sigma_\delta^2(0)$ where $\sigma_\delta^2(0)$ is the model error variance estimate when no explanatory variables are employed, but sampling errors are correctly deducted. Then, a pseudo coefficient of determination describing the fraction of the expected variability in the true y_i values that is explained by the model equals

$$R_\delta^2 = \frac{n[\sigma_\delta^2(0) - \sigma_\delta^2(k)]}{n\sigma_\delta^2(0)} = 1 - \frac{\sigma_\delta^2(k)}{\sigma_\delta^2(0)} \tag{2.20}$$

In actual practice the Bayesian mean values of $\sigma_\delta^2(k)$ and $\sigma_\delta^2(0)$ are employed. R_δ^2 is a direct extension of $\bar{R}^2$ in that it uses the ratio of unbiased estimators of the variance of the error δ and the variance of y. However, a critical difference is that $\bar{R}^2$ is based upon the sample variance of the observed y_i and the computed residual error $\hat{\varepsilon}_i$. Instead, R_δ^2 is based upon the estimated variance of the unobserved y_i and of the unobserved δ_i values. If $\hat{\sigma}_\delta^2(k) = 0$, this then yields the desired result $R_\delta^2 = 1$.

If no explanatory variables are employed, the R_δ^2 will be zero, as is expected. R_δ^2 fairly compares different WLS or different GLS models with varying numbers of parameters using the regional mean model ($k = 0$) as the base case.

Table 2.1 presents a pseudo Analysis of Variance (ANOVA) table for WLS or GLS. This table describes how much of the variation in the observations can be attributed to the model, the model error, and the sampling error, respectively. The total sampling error sum of squares can be described by its mean value, which is $tr[\mathbf{\Sigma}(\hat{\mathbf{y}})]$, where $tr[\mathbf{A}]$ is the trace of the matrix **A**. As noted above, because there are n observations, the total variation due to the model error δ for a model with k parameters has a mean equal to $n\sigma_\delta^2(k)$. This is called a pseudo ANOVA because the contributions of the three sources of error are estimated or constructed, rather than being determined from the computed residual errors and the observed model predictions. The impact of correlation among the sampling errors is ignored.

Table 2.1: Pseudo-ANOVA table for WLS and GLS regression analyses

Source	Degrees-of-Freedom	Sum of Squares
Model	k	$n\left[\sigma_\delta^2(0)-\sigma_\delta^2(k)\right]$
Model Error, δ_i	n-k-1	$n\sigma_\delta^2(k)$
Sampling Error, η_i	n	$tr\left[\sum(\hat{y})\right]$
Total	2n-1	$n\sigma_\delta^2(k)+tr\left[\sum(\hat{y})\right]$

EVR = $tr\left[\sum(\hat{y})\right]/n\sigma_\delta^2(k)$

MBV = $\frac{1}{n}w^T\Lambda(\sigma_\delta^2)w$ where w is the vector $\left(1/\sqrt{\Lambda(\sigma_\delta^2)_{ii}}\right)$

2.2.6 Error Variance Ratio and Misrepresentation of the Beta Variance

If the sampling variance is not small in comparison to the model error variance, a WLS or GLS analysis is more appropriate than an OLS analysis. The Error Variance Ratio (EVR) provides a measure of the relative importance of the sampling error compared to the model error. Thus, it provides an indication of the need for a WLS or GLS analysis. From Table 2.1, Griffis and Stedinger [2007] define EVR as

$$\text{EVR} = \frac{\text{SS(sampling error)}}{\text{SS(model error)}} = \frac{tr[\boldsymbol{\Sigma}(\hat{\mathbf{y}})]}{n\sigma_\delta^2} \tag{2.21}$$

As a rule of thumb, when the EVR is greater than 20%, one should employ a WLS or GLS analysis as opposed to OLS. If the EVR is less than 10%, the OLS results should be close to the WLS or GLS results depending upon the heterogeneity of the errors.

Although EVR distinguishes between the need for an OLS versus a WLS/GLS analysis, it does not determine whether a GLS regression is needed to address cross-correlation. Thus, the Misrepresentation of the Beta Variance (MBV) statistic was developed to determine whether a WLS regression is sufficient, or if a GLS regression is needed [Griffis and Stedinger, 2007]. The MBV describes the error made by a WLS

regression analysis in its evaluations of the precision of b_0^{WLS}, which is the estimator of the constant β_0. Covariance among the estimated y_i's generally has its greatest impact on the precision of the constant term [Stedinger and Tasker, 1985] and zero-one regional indicator variables. The corresponding values of the MBV is

$$\text{MBV} = \frac{Var\left[b_0^{WLS} \mid \text{GLS analysis}\right]}{Var\left[b_0^{WLS} \mid \text{WLS analysis}\right]} = \frac{w^T \Lambda w}{n} \quad \text{where } w_i = \frac{1}{\sqrt{\Lambda_{ii}}} \qquad (2.22)$$

If MBV is substantially larger than 1, then the GLS estimate of the variance of the constant term will be that much larger than the value provided by WLS. If all sites in the region have the same record length n, all records are concurrent, and all cross-correlations among the $\hat{y}_i$ are equal to ρ_η, MBV would be

$$\text{MBV} = 1 + (n-1)\rho_\eta \frac{\text{EVR}}{\text{EVR}+1} \qquad (2.23)$$

This special case shows the critical importance of the number of sites n in an analysis and the cross-correlation ρ_η of the at-site estimators $\hat{y}_i$. This indicates that for a B-GLS regional skew regression, wherein EVR = $Var(\hat{y}) / \sigma_\delta^2$ is generally greater than 2 or 3, the precision of the constant estimator will be particularly sensitive to cross-correlation ρ_η, and GLS is likely to be required to obtain statistically valid results.

2.2.7 Leverage and Influence

Leverage, as adapted by Tasker and Stedinger [1989, eqn. 23] considers whether an observation, or x-value, is unusual, and thus likely to have a large effect on the estimated regression coefficients [Cook and Weisberg, 1982]. How to measure leverage can be problematic. It is not clear how to describe how large a change in different residuals should be considered when model errors are heteroscedastic. This

leverage measure, suggested by Tasker and Stedinger [1989, eqn. 23], considers the effect of a unit change in each residual. If all the residuals have the same units and precision, then this is an appropriate measure of the effect of equivalent errors in the different observations. Thus, this leverage measures the marginal/unit impact of the residuals ε_i on the estimated y_i-values. In a Bayesian context, the leverage measure described by Tasker and Stedinger [1989] becomes

$$\text{leverage}(i) = \frac{\partial \hat{y}_i}{\partial \varepsilon_i} = \underset{\sigma_\delta^2}{E}\left[\mathbf{x}_i (\mathbf{X}^T \mathbf{\Lambda}\left(\sigma_\delta^2\right)^{-1} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{\Lambda}\left(\sigma_\delta^2\right)^{-1} e_i \right] \tag{2.24}$$

where $\mathbf{x}_i\mathbf{b}$ is the estimator of y_i associated with $\mathbf{x}_i$ and e_i is a unit column vector with 1 at the i-th row and zero otherwise Reis [2005]. Tasker and Stedinger [1989] show that the average value of this leverage statistic is $(k+1)/n$. Generally $2(k+1)/n$ is considered to be a large value.

This leverage statistic seems appropriate when lack-of-fit can be described by errors of the same magnitude. For example, this would be appropriate when modeling skew coefficients or the logarithm of the 100-year flood at different sites, as opposed to when some equations describe head and others flow in a groundwater model [Yager, 1998].

A second measure of leverage is Statistical leverage (S-leverage) which considers not a unit change in each residual, but a change proportional to the standard deviation of that residual [Reis, 2005]. Thus, this measure considers the likely statistical variation in each ε_i and the effect of such variation. S-leverage is computed as

$$\text{S-leverage}(i) = \omega(k+1)\frac{\partial \hat{y}_i}{\partial \varepsilon_i}\sigma_{\varepsilon_i} = \omega(k+1)\left(\text{leverage}(i)\cdot\sqrt{\Lambda\left(\sigma_\delta^2\right)_{ii}}\right),$$

$$\text{where } \omega = 1\Big/\sum_{j=1}^{n}\text{leverage}(i)\cdot\sqrt{\Lambda\left(\sigma_\delta^2\right)_{jj}} \tag{2.25}$$

As it is defined, the average value of S-leverage is also equal to $(k+1)/n$. Twice the average value, $2(k+1)/n$, is considered to be a large value.

S-leverage is an appropriate statistic to consider when the concern is with the likely effect on the regression of probabilistic variation in each residual. The GLS weights depend upon the statistical precision of each ε_i. Thus, the leverage in (2.24) for a point often increases as the at-site record length increases because of the greater weight assigned to the observation, whereas S-leverage in (2.25) is less dependent on record length. If an observation has no leverage, then given its anticipated statistical precision and the leverage associated with the corresponding **x**, the observation is unlikely to have any effect on estimated model parameters. The leverage in (2.24) may be more appropriate when one is concerned with the impact of gross errors in a model's structure, but it does not correct for differences in units among the ε_i. Examples in Chapter 3 and 4 illustrate the use of these statistics.

A third measure of leverage has been developed for use with Region-of-Influence (ROI) regression, in which a unique region, or set of gauged basins, is defined for each ungauged basin and is then used to predict hydrologic quantities such as flood quanitles [Eng *et al.*, 2007b]. The ROI leverage is computed as [Eng *et al*, 2007b]

$$\text{ROI-leverage}(i) = \frac{\partial \hat{y}_0}{\partial \varepsilon_i} = \left[\mathbf{x}_0 \left(\mathbf{X}^T \mathbf{\Lambda} \left(\sigma_\delta^2 \right)^{-1} \mathbf{X} \right)^{-1} \mathbf{X}^T \mathbf{\Lambda} \left(\sigma_\delta^2 \right)^{-1} \right] e_i \qquad (2.26)$$

where $\mathbf{x}_0$ is a vector of basin attributes at the ungauged sites. ROI leverage measures the impact on the estimate of y_0 at site 0 with $\mathbf{x} = \mathbf{x}_0$ of a unit error e_i at other sites i, i = 1 … n. In an ROI regression, one would like all the ROI-leverage statistics to be positive with approximately the same value.

Unlike leverage which highlights points which are likely to affect the fit of the regression, influence describes those points which do have an unusual impact on the regression analysis. An influential observation is one with an unusually large residual that has a disproportionate effect on the fitted regression relationships. Influential observations often have high leverage. The following influence measure, D_i, proposed by Tasker and Stedinger [1989] is based on Cook's D [Cook and Weisberg, 1982; Clarke, 1994],

$$D_i = \frac{k_{ii}\hat{\varepsilon}_i^2}{(k+1)(\lambda_{ii} - k_{ii})^2} \tag{2.27}$$

where $(k + 1)$ is the dimension of $\boldsymbol{\beta}$, λ_{ii} are the diagonal elements of $\boldsymbol{\Lambda}$, and k_{ii} are the diagonal elements of

$$\mathbf{K} = \mathbf{X}(\mathbf{X}^T\boldsymbol{\Lambda}^{-1}\mathbf{X})^{-1}\mathbf{X}^T \tag{2.28a}$$

so that $\lambda_{ii} - k_{ii}$ is the variance of $\hat{\varepsilon}_i$, as demonstrated in the Appendix. Tasker and Stedinger [1989] suggested influence is large when D_i is greater than $4/n$ where n is the number of sites. In a Bayesian analysis, as described in Reis [2005], one should employ $\mathbf{K_B}$, the average value of (2.27),

$$\mathbf{K_B} = \mathbf{X}\,\underset{\sigma_\delta^2}{E}\left[(\mathbf{X}^T\boldsymbol{\Lambda}\left(\sigma_\delta^2\right)^{-1}\mathbf{X})^{-1}\right]\mathbf{X}^T \tag{2.28b}$$

A second measure of influence, σ-influence [Gruber *et al.* 2007], determines which, if any, observations have an unusual impact on the estimated model error variance. In using regional skew models, the model error variance is very important because it determines the weight placed on the regional skew relative to the at-site estimator. The σ-influence statistic describes the relative influence of each observation on the estimated model error variance. The influence statistic D_i described in equation

(2.27) identifies those observations with significant influence on the model predictions. D_i does not necessarily describe whether the point has a significant influence on the estimated model error variance. The σ-influence is calculated as,

$$\sigma - i\text{nfluence}_i = \frac{2\sum_{j=1}^{n} \hat{\varepsilon}_i \left(\mathbf{\Lambda} \left(\sigma_\delta^2 \right)^{-1} \right)_{ij} \hat{\varepsilon}_j}{\sum_{i=1}^{n} \sum_{j=1}^{n} \hat{\varepsilon}_i \left(\mathbf{\Lambda} \left(\sigma_\delta^2 \right)^{-1} \right)_{ij} \hat{\varepsilon}_j} = \frac{2\hat{\varepsilon}_i \left(\mathbf{\Lambda} \left(\sigma_\delta^2 \right)^{-1} \hat{\varepsilon} \right)_i}{\hat{\varepsilon}^T \mathbf{\Lambda} \left(\sigma_\delta^2 \right)^{-1} \hat{\varepsilon}} \qquad (2.29)$$

Here the standardize sum-of-squares $\hat{\boldsymbol{\varepsilon}}^T \mathbf{\Lambda} \left(\sigma_\delta^2 \right)^{-1} \hat{\boldsymbol{\varepsilon}}$ used to compute the likelihood function for the data, and the generalized method of moments model error variance in Stedinger and Tasker [1985, 1986a], is divided among the *n* different sites. By construction, the average value of σ-influence is 2/*n*, where *n* is the number of sites in the regression; thus, σ-influence values greater than 4/*n* are considered to be large, as is the case with D_i.

2.3 Conclusion

The quasi-analytic Bayesian analysis of a GLS regression model described by Reis *et al*. [2005] has been developed into an operational GLS regional hydrologic regression methodology. Model selection criteria for use in conjunction with the Bayesian GLS regional hydrologic regression include the Average Variance of Prediction (AVP), as well as the Bayesian Plausibility Value ψ. Regression diagnostic statistics for WLS and GLS models include pseudo Analysis of Variance (pseudo ANOVA) tables a pseudo adjusted R_δ^2, Error Variance Ratio (EVR) and Misrepresentation of the Beta Variance (MBV), leverage and influence, and σ-influence.

APPENDIX A

VARIANCE OF THE RESIDUALS AND VARIANCE OF PREDICTION

This appendix provides a clean and consistent derivation of key expressions employed to compute the AVP for both new and old sites, as well as leverage and influence.

Variance of the Residuals

Let **y** be the vector of the true value of the statistic of interest, $\hat{\mathbf{y}}$ the at-site estimate, and $\hat{y}_p$ the prediction of y given by a fitted regression model. Then, the residual vector is given by

$$\begin{aligned}\hat{\boldsymbol{\varepsilon}} &= \hat{\mathbf{y}} - \mathbf{y}_p = \hat{\mathbf{y}} - \mathbf{X}\mathbf{b} \\ &= \hat{\mathbf{y}} - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\Lambda}\left(\sigma_\delta^2\right)^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\boldsymbol{\Lambda}\left(\sigma_\delta^2\right)^{-1}\hat{\mathbf{y}} = (\mathbf{I}-\mathbf{H})\hat{\mathbf{y}}\end{aligned} \tag{A.1}$$

wherein the GLS hat matrix **H** is defined as

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^{T}\boldsymbol{\Lambda}\left(\sigma_\delta^2\right)^{-1}\mathbf{X})^{-1}\mathbf{X}^{T}\boldsymbol{\Lambda}\left(\sigma_\delta^2\right)^{-1} \tag{A.2}$$

That the leverages in eqn. (24) are the average values of the diagonal elements h_{ii} of **H** follows from $\hat{\mathbf{y}}_p = \mathbf{H}\hat{\mathbf{y}}$. The average of the n leverage-values $tr[\mathbf{H}]/n$ equals $(k+1)/n$ because

$$\begin{aligned}\mathrm{tr}\left[\mathbf{H}\right] &= \mathrm{tr}[\mathbf{X}(\mathbf{X}^{T}\boldsymbol{\Lambda}\left(\sigma_\delta^2\right)^{-1}\mathbf{X})^{-1}\mathbf{X}^{T}\boldsymbol{\Lambda}\left(\sigma_\delta^2\right)^{-1}] \\ &= \mathrm{tr}[(\mathbf{X}^{T}\boldsymbol{\Lambda}\left(\sigma_\delta^2\right)^{-1}\mathbf{X})^{-1}\mathbf{X}^{T}\boldsymbol{\Lambda}\left(\sigma_\delta^2\right)^{-1}\mathbf{X}] = \mathrm{tr}[\mathbf{I}_{k+1}] = (k+1)\end{aligned} \tag{A.3}$$

For the OLS case, the hat matrix $\mathbf{H}$ has simply the traditional value $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$.

Substituting the equality $\hat{\mathbf{y}} = \mathbf{Xb} + \boldsymbol{\varepsilon}$ into (A.1), the computed residuals can be written as a function of the total error $\boldsymbol{\varepsilon}$:

$$\hat{\boldsymbol{\varepsilon}} = (\mathbf{I} - \mathbf{H})\boldsymbol{\varepsilon} \tag{A.4}$$

Given $E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T] = \boldsymbol{\Lambda}\left(\sigma_\delta^2\right)$, one finds that $\mathbf{H}\boldsymbol{\Lambda}\left(\sigma_\delta^2\right)\mathbf{H}^T = \boldsymbol{\Lambda}\left(\sigma_\delta^2\right)\mathbf{H}^T$ as can be demonstrated by substitution for $\mathbf{H}$ the expression in (A.2). Thus, the covariance matrix of the estimated residuals is

$$\begin{aligned} E[\hat{\boldsymbol{\varepsilon}}\hat{\boldsymbol{\varepsilon}}^T] &= E[(\mathbf{I} - \mathbf{H})\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T(\mathbf{I} - \mathbf{H}^T)] \\ &= E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T - \mathbf{H}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T - \boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T\mathbf{H}^T + \mathbf{H}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T\mathbf{H}^T] \\ &= \boldsymbol{\Lambda}\left(\sigma_\delta^2\right) - \mathbf{H}\boldsymbol{\Lambda}\left(\sigma_\delta^2\right) - \boldsymbol{\Lambda}\left(\sigma_\delta^2\right)\mathbf{H}^T + \mathbf{H}\boldsymbol{\Lambda}\left(\sigma_\delta^2\right)\mathbf{H}^T \\ &= (\mathbf{I} - \mathbf{H})\boldsymbol{\Lambda}\left(\sigma_\delta^2\right) \\ &= \boldsymbol{\Lambda}\left(\sigma_\delta^2\right) - \mathbf{K} \end{aligned} \tag{A.5}$$

which is clearly symmetric, and where $\mathbf{K}$ is defined in eqn. (28). For every σ_δ^2, the mean of $\hat{\boldsymbol{\varepsilon}}$ is zero, so in a Bayesian framework, the covariance of $\hat{\boldsymbol{\varepsilon}}$ is $E[\hat{\boldsymbol{\varepsilon}}\hat{\boldsymbol{\varepsilon}}^T] = E\left[\boldsymbol{\Lambda}\left(\sigma_\delta^2\right) - \mathbf{K}\right]$.

Similarly, the variance of $\boldsymbol{\beta}$ for a given $\boldsymbol{\Lambda}\left(\sigma_\delta^2\right)$ is obtained by noting that

$$\begin{aligned} \mathbf{b} &= (\mathbf{X}^T\boldsymbol{\Lambda}\left(\sigma_\delta^2\right)^{-1}\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Lambda}\left(\sigma_\delta^2\right)^{-1}\hat{\mathbf{y}} \\ &= (\mathbf{X}^T\boldsymbol{\Lambda}\left(\sigma_\delta^2\right)^{-1}\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Lambda}\left(\sigma_\delta^2\right)^{-1}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) \\ &= \boldsymbol{\beta} + (\mathbf{X}^T\boldsymbol{\Lambda}\left(\sigma_\delta^2\right)^{-1}\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Lambda}\left(\sigma_\delta^2\right)^{-1}\boldsymbol{\varepsilon} \end{aligned} \tag{A.6}$$

Thus, for given $\boldsymbol{\Lambda}\left(\sigma_\delta^2\right)$

$$E[(\boldsymbol{\beta}-\mathbf{b})(\boldsymbol{\beta}-\mathbf{b})^T] = E[(\mathbf{X}^T\boldsymbol{\Lambda}\left(\sigma_\delta^2\right)^{-1}\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Lambda}\left(\sigma_\delta^2\right)^{-1}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T\boldsymbol{\Lambda}\left(\sigma_\delta^2\right)^{-1}\mathbf{X}(\mathbf{X}^T\boldsymbol{\Lambda}\left(\sigma_\delta^2\right)^{-1}\mathbf{X})^{-1}]$$
$$= (\mathbf{X}^T\boldsymbol{\Lambda}\left(\sigma_\delta^2\right)^{-1}\mathbf{X})^{-1} \quad \text{(A.7)}$$

Variance of Prediction

The vector of differences between the true and predicted values is

$$\begin{aligned}(\mathbf{y}-\mathbf{y}_p) &= \mathbf{X}\boldsymbol{\beta}+\boldsymbol{\delta}-\mathbf{X}\mathbf{b}\\ &= \mathbf{X}\boldsymbol{\beta}+\boldsymbol{\delta}-\mathbf{X}(\mathbf{X}^T\boldsymbol{\Lambda}\left(\sigma_\delta^2\right)^{-1}\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Lambda}\left(\sigma_\delta^2\right)^{-1}(\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}) \quad \text{(A.8)}\\ &= \boldsymbol{\delta}-\mathbf{H}\boldsymbol{\varepsilon}\end{aligned}$$

Thus, the covariance matrix for the prediction errors for given Λ is just

$$\begin{aligned}E[(\mathbf{y}-\mathbf{y}_p)(\mathbf{y}-\mathbf{y}_p)^T] &= E[(\boldsymbol{\delta}-\mathbf{H}\boldsymbol{\varepsilon})(\boldsymbol{\delta}-\mathbf{H}\boldsymbol{\varepsilon})^T]\\ &= E[\boldsymbol{\delta}\boldsymbol{\delta}^T-\boldsymbol{\delta}\boldsymbol{\varepsilon}^T\mathbf{H}^T-\mathbf{H}\boldsymbol{\varepsilon}\boldsymbol{\delta}^T+\mathbf{H}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T\mathbf{H}^T] \quad \text{(A.9)}\end{aligned}$$

Noting that $\boldsymbol{\delta}$ for new sites and $\boldsymbol{\varepsilon}$ for old sites are uncorrelated, the covariance matrix for the predictions at new sites is simply

$$E[(\mathbf{y}-\mathbf{y}_p)(\mathbf{y}-\mathbf{y}_p)^T] = E[\boldsymbol{\delta}\boldsymbol{\delta}^T+\mathbf{H}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T\mathbf{H}^T] \quad \text{(A.10)}$$

Therefore, substituting $E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T] = \Lambda$ with H from (A.2), the variance of prediction at a new site is given by

$$VP_{new}(i) = \sigma_\delta^2 + \mathbf{x}_i(\mathbf{X}^T\boldsymbol{\Lambda}\left(\sigma_\delta^2\right)^{-1}\mathbf{X})^{-1}\mathbf{x}_i^T \quad \text{(A.11)}$$

However, if the predictions are made for those *n* old sites used in the regression, the covariance matrix of the predictions becomes

$$E[(\mathbf{y}-\mathbf{y}_p)(\mathbf{y}-\mathbf{y}_p)^T] = E[\boldsymbol{\delta}\boldsymbol{\delta}^T - 2\boldsymbol{\delta}\boldsymbol{\delta}^T\mathbf{H}^T + \mathbf{H}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T\mathbf{H}^T] \\ = \sigma_\delta^2\mathbf{I} + \mathbf{H}\boldsymbol{\Lambda}\left(\sigma_\delta^2\right)\mathbf{H}^T - 2\sigma_\delta^2\mathbf{H}^T \tag{A.12}$$

because the model error δ_i for the site is also part of the sampling error of the estimator. Thus, the variance of prediction for an old site is

$$VP_{old}(i) = \sigma_\delta^2 + \mathbf{x}_i(\mathbf{X}^T\boldsymbol{\Lambda}\left(\sigma_\delta^2\right)^{-1}\mathbf{X})^{-1}\mathbf{x}_i^T - 2\sigma_\delta^2\mathbf{x}_i(\mathbf{X}^T\boldsymbol{\Lambda}\left(\sigma_\delta^2\right)^{-1}\mathbf{X})^{-1}\mathbf{X}^T e_i \tag{A.13}$$

wherein e_i is a column vector with one at the i^{th} row and zero otherwise. In a Bayesian analysis wherein σ_δ^2 is a random variable, one should employ the appropriate expected values in eqn. (A.2), (A.7), (A.11), and (A.13), as in eqn. (12)-(17).

REFERENCES

Bayarri, M.J. and Berger, J.O. (2000), P-values for composite null models. *Journal of the American Statistical Association* 95, 1127-1142.

Bickel, P. and Doksum, K., (1977), *Mathematical Statistics: Basic Ideas and Selected Topics,* Holden-Day, San Francisco.

Carlin B.P. and T.A. Louis, (2000), *Bayes and Empirical Bayes Methods for Data Analysis,* Chapman & Hall/CRC, Boca Raton, FL.

Clarke, R. T., (1994), *Statistical Modeling in Hydrology*, John Wiley & Sons Inc.

Cook, R.D. and Weisberg, S., (1982), Residuals and Influence in Regression, Chapman and Hall, New York, NY, 230 pp.

Devore, Jay L. (2004), *Probability and Statistics: For Engineering and the Sciences*, Thomson, United States, 795 pp.

Eng, K., Stedinger, J.R., and Gruber, A.M., (2007b), Regionalization of streamflow characteristics for the Gulf-Atlantic Rolling Plains using leverage guided region-of -influence regression, in Kabbes, K.C., ed., Proceedings of the World Environmental and Water Resources Congress, May 15–19, 2007, Tampa, Florida, USA: American Society of Civil Engineers.

Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B., (2004), *Bayesian Data Analysis*, Chapman & Hall/CRC, Boca Raton, FL.

Griffis, V. W., and J. R. Stedinger, (2007), The Use of GLS Regression in Regional Hydrologic Analyses, J. of Hydrology, 344(1-2), 82-95, [doi:10.1016/j.jhydrol.2007.06.023].

Gruber, Andrea M., Dirceu S. Reis Jr., and Jery R. Stedinger, (2007), Models of Regional Skew Based on Bayesian GLS Regression, Paper 40927-3285, World Environmental & Water Resources Conference - Restoring our Natural Habitat, K.C. Kabbes editor, Tampa, Florida, May 15-18.

Han, S., S. Guikema, S. Quiring, K. Lee, D. Rosowsky, and R. Davidson, (2009), Estimating the spatial distribution of power outages during hurricanes in the gulf coast region. *Reliability Engineering & System Safety 94* (2), 199-210.

Kitanidis, P. K., (1986), "Parameter Uncertainty in Estimation of Spatial Functions: Bayesian Analysis," *Water Resources Research*, 22 (4), 499-507.

Linart, H. and W. Zucchini, (1986), *Model Selection*, John Wiley and Sons, Inc., New York.

Lindley, D.V., (1965), *Introduction to Probability and Statistics from a Bayesian Viewpoint, Part2. Inference*. Cambridge: University Press.

Liu, H., Davidson, R. A., Rosowsky, D. V., Stedinger, J. R. (2005), Negative binomial regression of electric power outages in hurricanes. ASCE Journal of Infrastructure Systems , 11 (4), 258–267.

Madsen, H., and D. Rosbjerg, (1997), Generalized least squares and empirical Bayes estimation in regional partial duration series index-flood modeling, *Water Resour Res., 33*(4), 771-782.

Madsen, H., P. S. Mikkelsen, D. Rosbjerg, and P. Harremoe¨s, (2002), Regional estimation of rainfall intensity-duration-frequency curves using generalized least squares regression of partial duration series statistics, Water Resour. Res., 38(11), 1239, doi:10.1029/2001WR001125.

Qian, S. S.; Reckhow, K. H.; Zhai, J., McMahon, G., (2005), Nonlinear regression modeling of nutrient loads in streams: A Bayesian approach, *Water Resour. Res., 41*(7), W07012*,* doi: 10.1029/2005WR003986.

Reis Jr., D.S., (2005). Flood Frequency Analysis Employing Bayesian Regional Regression and Imperfect Historical Information. Ph.D. Dissertation, Cornell University.

Reis, D. S., Jr., J. R. Stedinger, and E. S. Martins, (2005), Bayesian generalized least squares regression with application to log Pearson type 3 regional skew estimation, *Water Resour. Res.*, *41*, W10419, doi:10.1029/2004WR003445.

Rencher, A. C., (2000), *Linear Models in Statistics*, Wiley Series in Probability and Statistics, John Wiley & Sons, Inc.

Robins, J.M., van der Vaart, A., and Ventura, V. (2000), The asymptotic distribution of *p*-values in composite null models. *Journal of the American Statistical Association*, 95, 1143-1156.

Stedinger, J.R., and G.D. Tasker, (1985), Regional Hydrologic Analysis, 1. Ordinary, Weighted and Generalized Least Squares Compared, *Water Resources Research, 21*(9), 1421-1432.

Stedinger, J.R. and G. Tasker, (1986a), Correction to "Regional hydrologic analysis, 1, Ordinary, weighted and generalized least squares compared", *Water Res. Research*, 22(5), 844, 1986.

Stedinger, J.R. and G. Tasker, (1986b), Regional hydrologic analysis, 2: Model-error estimators, estimation of sigma and log-Pearson Type 3 distributions, *Water Res. Research*, 22(10), 1487-1499, 1986b.

Tasker, G.D., and J.R. Stedinger, (1986), Regional skew with weighted least squares regression, *Journal of Water Resources Planning and Management*, 112(2), 225-237.

Tasker, G.D., and J.R. Stedinger, (1989), An Operational GLS Model for Hydrologic Regression, *Journal of Hydrology, 111*(1-4), 361-375.

Yager, R. M., (1998), Detecting Influential Observations in Nonlinear Regression Modeling of Groundwater Flow, *Water Resources Research*, 34(7), 1623-1633.

Zellner, A., (1971), *An Introduction to Bayesian Inference in Econometrics*, John Wiley and Sons, Inc., New York.

CHAPTER 3

APPLICATION OF BAYESIAN-GENERALIZED LEAST SQUARES FRAMEWORK TO REGIONAL SKEW REGRESSION

3.1 Application of Regional Skew Estimation to Illinois River Basin Data Set and the State of South Carolina Data Set

Bulletin 17B [IACWD, 1982] recommends the use of the log-Pearson Type III distribution for flood frequency analysis. The data available at a given site are usually too short to provide a good estimate of the skewness coefficient. In order to improve the precision of the skewness estimator, *Bulletin 17B* recommends combining a regional skew with the at-site skew [Hardison, 1975; Tasker, 1978; McCuen, 1979 and 2001; IACWD, 1982; Stedinger *et al*., 1993; Griffis and Stedinger, 2004]. McCuen and Smith [2008] recently developed a regression model for the skewness coefficient based on rainfall skew and basin storage variables. However, it is unclear whether they are addressing log-space or real space skewness. Reis *et al*. [2005] evaluated regional skew for 17 sites in the Tibagi River basin and for 44 sites in the Muskingum River basin. For this study, regional skew was evaluated for two different and larger data sets: the Illinois River basin with 62 sites with record lengths ranging from 14 to 90 years, and the state of South Carolina with 89 stations with record lengths varying from 25 to 104 years. They compared the results of OLS, WLS, and GLS regional regression analyses using method of moments (MM) and Bayesian model-error-variance estimators [Reis, 2005].

Binary variables were used to verify if any variability in the at-site skews could be explained by hydrologic region. The analysis for the State of South Carolina included sites from North Carolina and Georgia, and the whole area was divided into

four hydrologic regions with three binary variables (Z_1, Z_2, Z_3) as follows: Blue Ridge (0, 0, 0), Piedmont (1, 0, 0), Upper Coastal Plain (0, 1, 0), and Lower Coastal Plain (0, 0, 1) [Feaster and Tasker, 2002]. The Illinois River basin was divided into three regions, as described in Tasker and Stedinger [1986]. The regions and values of the binary variables (Z_1, Z_2) for each were: Little Wabash (1, 0), Rock (0, 1), and Sangamon (0, 0).

All explanatory variables, except the binary variable, were centered by subtracting their means so that the constant and the binary variables could be used to compute the regional mean of each hydrologic region. The logarithms of the following basin characteristics were used in the regression models and all possible combinations of these variables were tested.

South Carolina: (1) Drainage area, A, in square miles; (2) main channel slope in feet per mile; (3) length of the main channel in miles; (4) the mean basin elevation in feet above sea level; (5) annual mean precipitation in inches at the centroid of the basin; (6) annual mean runoff in inches at the centroid of the basin.

Illinois: (1) Drainage area, A, in square miles; (2) main channel slope in feet per mile; (3) area of lakes, expressed as percentage of drainage area and increased by one; (4) forest cover expressed as percentage of drainage area and increased by one; and (5) soil permeability index which varies from (1)low infiltration to (6) high infiltration.

3.2 Sampling Covariance Matrix

Estimates of $\sigma^2_{\eta i}$ and $\hat{\rho}(\hat{y}_i, \hat{y}_j)$ are required. Griffis and Stedinger [2004] provide an accurate approximation of $\sigma^2_{\eta_i}$ that, with the skewness estimator unbiasing factor in Stedinger and Tasker [1986b], equals:

$$Var(G) = \left[1 + \frac{6}{m}\right]^2 \left(\frac{6}{m} + a(m)\right)\left(1 + \left\{\frac{9}{6} + b(m)\right\}\gamma^2 + \left\{\frac{15}{6*8} + c(m)\right\}\gamma^4\right) \qquad (3.1)$$

wherein

$$a(m) = -\frac{17.75}{m^2} + \frac{50.06}{m^3},$$

$$b(m) = \frac{3.92}{m^{0.3}} - \frac{31.1}{m^{0.6}} + \frac{34.86}{m^{0.9}},$$

and $$c(m) = -\frac{7.31}{m^{0.59}} + \frac{45.9}{m^{1.18}} - \frac{86.5}{m^{1.77}}$$

Here m is the sample size and γ is the true value of skew. The factor $[1 + 6/m]^2$ in equation (3.1) should be employed when the bias correction factor proposed by Tasker and Stedinger [1986] is used in the estimation of the at-site skew, computed as

$$G = \left[1 + \frac{6}{m}\right] \frac{m \sum_{t=1}^{m} (w_t - \overline{w})^3}{(m-1)(m-2)s^3} \qquad (3.2)$$

where w_t is the logarithm of the annual peak flows in year t, and s is the sample standard deviation of w_t. Because the true values of skews at each site are unknown, the regional mean of the skews is used in equation (3.1).

Martins and Stedinger [2002] express the inter-site correlation coefficient between two G_i in terms of the inter-site correlation coefficient ρ_{ij} between concurrent flows as

$$\hat{\rho}(G_i, G_j) = (cf_{ij})\rho_{ij}^{\kappa} \qquad cf_{ij} = m_{ij} \Big/ \sqrt{(m_{ij} + m_i)(m_{ij} + m_j)} \qquad (3.3)$$

where the exponent κ depends upon the regional value for γ, and is equal to 2.8 for $\gamma =$ 0, m_{ij} is the common record period, and m_i and m_j are the extra observation period for station i and j, respectively. The factor (cf_{ij}) accounts for the sample size differences between stations, and the concurrent record length: it worked reasonably well for the skew-coefficient in the range considered ($-1 \le \gamma \le 1$).

The use of sample estimates of ρ_{ij} may result in a covariance matrix $\mathbf{\Lambda}\left(\sigma_\delta^2\right)$ that is not positive definite due to sampling uncertainties and variations in concurrent record lengths [Tasker and Stedinger, 1989]. Therefore, one can use a smoothed estimate of ρ_{ij} that depends on the distances between any two stations $\rho(d_{ij})$. For the two examples,

$$\rho(d_{ij}) = \theta^{\left(\frac{\tau d_{ij}}{\alpha d_{ij}+1}\right)} \tag{3.4}$$

wherein $\theta = 0.988$, $\alpha = 0.002$, and $\tau = 3$ for the Illinois River basin; and $\theta = 0.990$, $\alpha =$ 0.0023, and $\tau = 2.8$ for the state of South Carolina; d_{ij} is the distance between sites in kilometers.

3.3 Regression Analysis for Illinois River Basin and State of South Carolina Data Sets

The Bayesian Generalized Least Squares (B-GLS) regression framework for hydrologic statistics addressed in Chapter 2 is applied here to develop a regional skewness estimator for both the Illinois River Basin data set and the state of South Carolina data set. Ordinary Least Squares (OLS), Weighted Least Squares (WLS), and Generalized Least Squares (GLS) regression analysis were also performed to compare to the B-GLS results. Tables 3.1 and 3.2, and Tables 3.3 and 3.4, present the results of the analysis for the state of South Carolina and the Illinois River basin, respectively. The following sections describe the results in detail.

Table 3.1: Regional skew regression results for the state of South Carolina data set (number of sites = 89). Table reports best models in terms of minimum average variance of prediction (AVP) for a new site. Standard errors, classical (OLS and MM) and Bayesian plausibility values (%) are presented in parentheses. Average regional Effective Record Length (ERL), as well as, Average Sampling Variance (ASV) is reported for each model.

Model	Regression Parameters				Regression Diagnostics				
	Const	Z_1	ln(S)	ln(L)	σ_δ^2	ASV	AVP_{new}	R^2_δ	ERL(yrs)
Ordinary Least Squares Regression Results									
MM-OLS	0.10 (0.045)	-0.26 (0.072) (0.053%)	-	-	0.11	0.0025	0.11	0.12	63
Weighted Least Squares Regression Results									
MM-WLS	-0.028 (0.041)	-	-	-	0.00	0.0017	0.0017	0.00	3014
B-WLS 1	-0.028 (0.044)	-	-	-	0.014 (0.012)	0.0017	0.015	0.00	409
B-WLS 2	0.081 (0.054)	-0.28 (0.082) (0.19%)	-	-	0.010 (0.010)	0.0034	0.013	0.26	486
Generalized Least Squares Regression Results									
MM-GLS 1	0.062 (0.091)	-0.21 (0.094) (3.0%)	-0.18 (0.041) (0.0%)	-0.20 (0.058) (0.13%)	0.00	0.012	0.013	0.99	481
B-GLS 1	0.076 (0.10)	-0.23 (0.10) (4.1%)	-0.14 (0.048) (2.0%)	-0.16 (0.061) (2.6%)	0.024 (0.013)	0.013	0.035	0.34	182
B-GLS Constant	0.0035 (0.090)	-	-	-	0.037 (0.013)	0.0069	0.043	0.00	148
B-GLS 2	0.0050 (0.091)	-	-0.13 (0.048) (2.4%)	-0.15 (0.061) (3.2%)	0.026 (0.013)	0.010	0.034	0.30	185
Sensitivity Analysis- Generalized Least Squares Regression Results									
B-GLS 2 (w/o site 81)	0.012 (0.089)	-	-0.057 (0.054) (32%)	-0.10 (0.063) (13%)	0.012 (0.010)	0.011	0.022	0.67	282
B-GLS 2 (w/o site 62)	0.013 (0.091)	-	-0.14 (0.047) (1.7%)	-0.19 (0.061) (1.2%)	0.025 (0.013)	0.010	0.034	0.32	185
B-GLS 2 (w/o site 53)	-0.012 (0.090)	-	-0.17 (0.043) (0.47%)	-0.19 (0.056) (0.68%)	0.019 (0.011)	0.010	0.028	0.49	223

Table 3.2: Pseudo ANOVA table for the state of South Carolina data set (B-GLS 2).

Source	Degrees-of-Freedom		Sum of squares			
	Case 1	Cases 2, 3, 4	Case 1 (all sites)	Case 2 (w/o site 53)	Case 3 (w/o site 62)	Case 4 (w/o site 81)
Model	k=2	k=2	1.0	1.6	1.0	2.2
Model Error	n-k-1=87	n-k-1=86	2.3	1.7	2.2	1.1
Sampling Error	n=89	n=88	16	16	15	16
Total	2n-1=177	2n-1=175	18	17	18	17
EVR			6.9	9.5	7.0	15
MBV			5.4	5.4	5.4	5.3
R^2_δ			0.30	0.49	0.32	0.67

Table 3.3: Regional skew regression results for the Illinois River data set (number of sites = 62). Table contains best models in terms of minimum average variance of prediction (AVP) for a new site. Standard errors, classical (OLS and MM) and Bayesian p-values (%) are presented in parentheses. Average regional Effective Record Length (ERL) is reported for each model.

Model	Regression Parameters				Regression Diagnostics				
	Const	Z_2	ln(A)	ln(S)	σ_δ^2	ASV	AVP_{new}	R^2_δ	ERL(yrs)
Ordinary Least Squares Regression Results									
MM-OLS	-0.077	-0.72	0.17	0.47	0.35	0.022	0.37	0.25	25
	(0.11)	(0.16)	(0.083)	(0.19)					
		(<0.01%)	(4.7%)	(1.6%)					
Weighted Least Squares Regression Results									
MM-WLS	-0.13	-0.52	-	0.12	0.064	0.016	0.080	0.55	101
	(0.10)	(0.14)		(0.05)					
		(<0.1%)		(2.7%)					
B-WLS	-0.13	-0.52	-	0.12	0.082	0.012	0.087	0.38	94
	(0.10)	(0.14)		(0.053)	(0.05)				
		(<0.1%)		(2.9%)					
Generalized Least Squares Regression Results									
MM-GLS 1	-0.092	-0.51	-	0.13	0.12	0.039	0.16	0.24	52
	(0.17)	(0.24)		(0.058)					
		(3.7%)		(2.6%)					
B-GLS 1	-0.09	-0.51	-	0.13	0.13	0.029	0.15	0.13	55
	(0.17)	(0.24)		(0.059)	(0.051)				
		(3.4%)		(2.7%)					
B-GLS constant	-0.42	-	-	-	0.15	0.010	0.16	0.00	52
	(0.12)				(0.056)				
B-GLS 2	-0.31	-	-	0.13	0.13	0.019	0.14	0.12	53
	(0.13)			(0.059)	(0.052)				
				(3.3%)					
Sensitivity Analysis- Generalized Least Squares Regression Results									
B-GLS 2 (w/o site 25)	-0.31	-	-	0.13	0.12	0.018	0.13	0.11	58
	(0.13)			(0.058)	(0.049)				
				(3.1%)					
B-GLS 2 (w/o site 9)	-0.29	-	-	0.12	0.12	0.019	0.13	0.11	57
	(0.13)			(0.058)	(0.051)				
				(4.9%)					
B-GLS 2 (w/o site 5)	-0.36	-	-	0.11	0.12	0.019	0.13	0.23	60
	(0.13)			(0.059)	(0.048)				
				(7.3%)					

Table 3.4: Pseudo ANOVA table for the Illinois River Basin data set (B-GLS 2).

Source	Degrees-of-Freedom		Sum of squares			
	Case 1	Cases 2, 3, 4	Case 1 (all sites)	Case 2 (w/o site 25)	Case 3 (w/o site 9)	Case 4 (w/o site 5)
Model	k=1	k=1	1.1	1.9	1.9	2.1
Model Error	n-k-1=60	n-k-1=59	8.2	7.3	7.2	7.0
Sampling Error	n=62	n=61	19	18	19	19
Total	2n-1=123	2n-1=121	27	26	26	26
EVR			2.3	2.5	2.6	2.7
MBV			4.2	4.1	4.1	4.2
R^2_δ			0.12	0.11	0.11	0.23

3.3.1 Ordinary Least Squares Regression Analysis

As expected, the method of moments OLS (MM-OLS) model error variance estimates for both the South Carolina and Illinois data sets are much larger than those obtained with WLS and GLS. In the case of South Carolina, the OLS model with minimum AVP for a new site is comprised of a constant and the binary variable Z_1, representing the Piedmont region. Thus, this model implies that the regional skew is a constant value in each of the four regions. As shown in Table 3.1, the model error variance, σ_δ^2, for MM-OLS is 0.11 with an AVP_{new} of 0.11. These values are much larger than any of the results obtained using WLS or GLS analysis.

Likewise, similar results can be observed for the Illinois data set. In this case, the model with minimum AVP for a new site contains a constant, the binary variable Z_2 representing the Rock region, the natural log of the basin main channel slope, and the natural log of the basin main channel length. As shown in Table 3.3, σ_δ^2, for MM-OLS is 0.35 with an AVP_{new} of 0.37. Again, these values are much larger than any of the results obtained using a WLS or GLS analysis.

The exaggerated variance of predictions occurs because the OLS regression analysis does not make any distinction between the variance due to the model error and the variance due to time sampling error. This concept is reinforced when viewing

the error variance ratio EVR results from the pseudo ANOVA tables, Tables 3.2 and 3.4, for both South Carolina and Illinois, respectively. The EVR for Case 1 for both South Carolina and Illinois are very large, 6.9 and 2.3, respectively, implying that the variation due to sampling error is much larger than the variation due to model error. Clearly, a WLS or GLS analysis should be employed rather than an OLS analysis.

3.3.2 Weighted Least Squares Regression Analysis

In the case of the weighted least squares regression, both the method of moments and Bayesian estimators were used. The best model when using the method of moments estimator for South Carolina is the constant model, which indicates that skew is equal to the same constant value of, -0.028 in all regions. This model MM-WLS, as presented in Table 3.1, has an AVP_{new} of 0.0017 and a model error variance equal to 0. This result is considered to be unrealistic as it underestimates the variance of prediction of the regional skew. Consequently, the weighted skew at any site, which is a variance-weighted average of the at-site and the regional skew estimates, will be too heavily weighted towards the regional value. For comparison, Table 3.1 also shows the results of the Bayesian constant model, B-WLS 1. These results appear more reasonable with an AVP_{new} of 0.015 and σ_δ^2 equal to 0.014. However, this was not the best B-WLS model for South Carolina. The model with the smallest AVP_{new}, is B-WLS 2, with the explanatory variable Z_1. This model has an AVP_{new} of 0.013 and σ_δ^2 is 0.010.

For the Illinois data set, the model with minimum AVP_{new} has a constant, the binary variable Z_2, and ln(Slope) as explanatory variables regardless of the model error variance estimator employed. As shown in Table 3.3, the estimated β-parameters obtained by MM-WLS and B-WLS are nearly identical, but both the AVP_{new} and model error variance are slightly smaller for the MM-WLS model as compared to the

B-WLS model. The larger value of AVP_{new} for the B-WLS model results in computed effective record lengths for regional skew estimates that are smaller than those based on the MM-WLS model.

In order to determine if a WLS analysis is sufficient, the misrepresentation of beta variance MBV can be consulted. As shown in Table 3.2, Case 1 for South Carolina has a MBV of 5.4. Table 3.4 shows that Case 1 for Illinois has an MBV of 4.2. These results clearly suggest, for both South Carolina and Illinois, that the correlation among estimators of skew in these regions should not be neglected by using a WLS analysis; otherwise the model error variance as well as the AVP of the regional skew will be underestimated.

3.3.3 Generalized Least Squares Regression Analysis

In the case of generalized least squares regression, both the method of moments and Bayesian estimators were used. For the South Carolina data set, the choice of the best model depends on the estimator employed. With the method of moments estimator, the best model (denoted MM-GLS 1) employs a constant, the binary variable Z_1, ln(Slope), and ln(Length) as explanatory variables. As shown in Table 3.1, MM-GLS 1 has a model error variance equal to 0 and an AVP_{new} equal to 0.01. As discussed previously, a zero model error variance is unrealistic.

The equivalent B-GLS model to the MM-GLS model was run for comparison. As shown in Table 3.1, B-GLS 1 has a more believable σ_δ^2 with a mean of 0.024 (and a standard deviation of 0.013) with an AVP_{new} of 0.035. However, the model with the minimum AVP_{new} for a new site when the Bayesian estimator is employed is B-GLS 2. This model's explanatory variables include a constant, ln(Slope), and ln(Length). The resultant σ_δ^2 for B-GLS 2 is 0.026 as compared to 0.024 for B-GLS 1, but the AVP_{new} for B-GLS 2 as compared to B-GLS 1 is about 0.001 smaller.

The choice of the best model for the Illinois data again depends on the estimator employed. When the method of moments estimator is used, the best model (denoted MM-GLS 1) employs a constant, the binary variable Z_2, and ln(Slope). The MM-GLS 1 model appears in Table 3.3, along with the corresponding Bayesian analysis, denoted B-GLS 1. As shown in Table 3.3, the estimated β-parameters obtained by MM-GLS 1 and B-GLS 1 are nearly identical, but the model error variance is slightly smaller for the MM-GLS 1 model as compared to the B-GLS 1 model. However, the AVP_{new} is slightly larger for the MM-GLS 1 model as compared to the B-GLS 1 model (AVP_{new} is equal to 0.16 and 0.15, respectively). The B-GLS 1 model is not the Bayesian GLS model with the minimum AVP_{new} for a new site, instead B-GLS 2 with a constant and ln(Slope) as its only explanatory variable has the smallest AVP_{new}. B-GLS 2 has an almost identical model error variance (0.13, with a standard deviation of 0.05) to B-GLS 1. However, the AVP_{new} for B-GLS 2 is 0.14 as compared to 0.15 for B-GLS 1, a modest difference.

3.3.4 Sensitivity Analysis and Diagnostic Statistics for B-GLS Analysis

This section offers a diagnostic analysis of the best B-GLS models from both the South Carolina and Illinois data. Figure 3.1 presents the leverage and influence results for the B-GLS 2 model from South Carolina.

The 22 sites included in Figure 3.1 have high leverage, high influence, high statistical leverage and/or high σ-influence. The sites are ordered by decreasing influence, as it identifies those sites that had a large impact on the analysis. Site 81 has the highest σ-influence value, implying it has a large impact on the model error variance, while also having high influence and high leverage values. Site 81 was removed from the B-GLS 2 analysis as a test. As expected, σ_{δ}^2 of B-GLS 2 (w/o site 81) decreased from 0.026 to 0.012, as shown in Table 3.1. When Site 81 is removed

from the regression both of the 95% credible regions for the regression coefficients of ln(Slope) and ln(Length) contain zero (Bayesian plausibility-values, ψ = 32% and 13%, respectively). This indicates that a simple regional mean model would be appropriate. Site 81 has such a large impact on the regression analysis because it has a long record length, 78 years, a large positive at-site skew value, 0.71, and the 6th largest residual, 0.66. When Site 81 is removed from the B-GLS 2 model, the R_s^2 increases from 0.30 to 0.67, which is a very large difference.

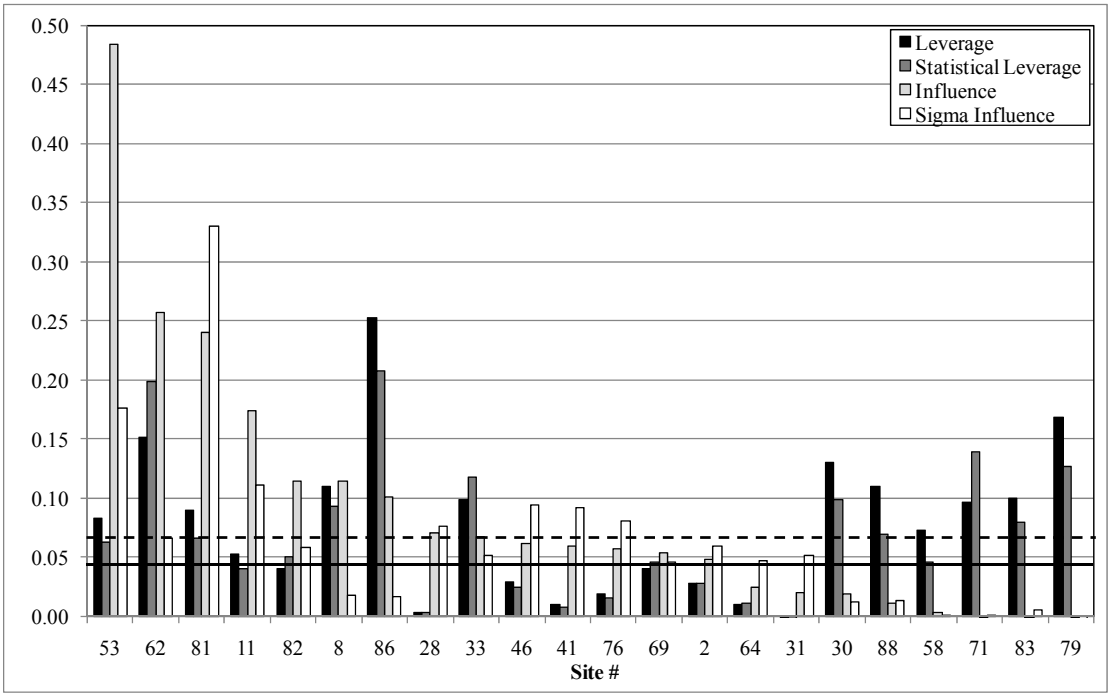

Figure 3.1: Regression Diagnostics: Leverage and influence for the state of South Carolina data set B-GLS 2 Model. The dashed, horizontal line represents the threshold for high leverage and high statistical leverage, while the solid, horizontal line represents the threshold for high influence and high σ-influence.

We also removed Site 62 from the B-GLS 2 model as a test. Like Site 81, Site 62 has a high and almost identical influence value, and has high leverage. However, the main difference can be seen in the σ-influence values; the σ-influence value of Site

81 is over 5 times larger than that of Site 62. As shown in Table 3.1, when Site 62 is removed, neither σ_δ^2 nor the AVP_{new} change from those generated for B-GLS 2. The small impact on the regression parameters is due to the fact that Site 62 has only 29 years of record. Site 62 does have an impact on the analysis because it has a large negative skew, -0.62, and the second largest residual in the study, -0.88.

The last sensitivity analysis run on the South Carolina data was the removal of Site 53, which had the largest influence in the study as well as high leverage and high σ-influence. As shown in Table 3.1, when Site 53 is removed from the B-GLS 2 model, there is a decrease in σ_δ^2 from 0.026 to 0.019, and the value of R_δ^2 increases from 0.30 to 0.49. Site 53 has a long record, 74 years, an at-site skew of 0.49 and the third largest residual in the study, 0.72.

Table 3.2 contains the pseudo-ANOVA results for both the B-GLS 2 model and the three sensitivity analysis. The pseudo-ANOVA table clearly demonstrates that for all four cases in South Carolina the sampling error is many times larger than the model error, EVR >7.

Figure 3.2 displays the leverage and influence results for the B-GLS 2 model for Illinois. The 30 sites included in Figure 3.2 have high leverage, high influence, high statistical leverage and/or high σ-influence. The sites are ordered by decreasing influence. Among the twenty-two sites that have high influence, only three also have high leverage. Site 25 is a site with a very large σ-influence and a moderately high influence. However, it has almost no leverage or statistical leverage. Site 25 has a record length of 25 years, a skew of -1.9, and a residual of -1.5, the second largest residual magnitude among the 62 sites used in the regression. When Site 25 is removed from the B-GLS 2 model as test, a decrease in the model error variance and the AVP_{new} is observed (σ_δ^2 is 0.13 in B-GLS 2 and it drops to 0.12 in B-GLS 2 w/o Site 25, while the AVP_{new} drops from 0.14 to 0.13.

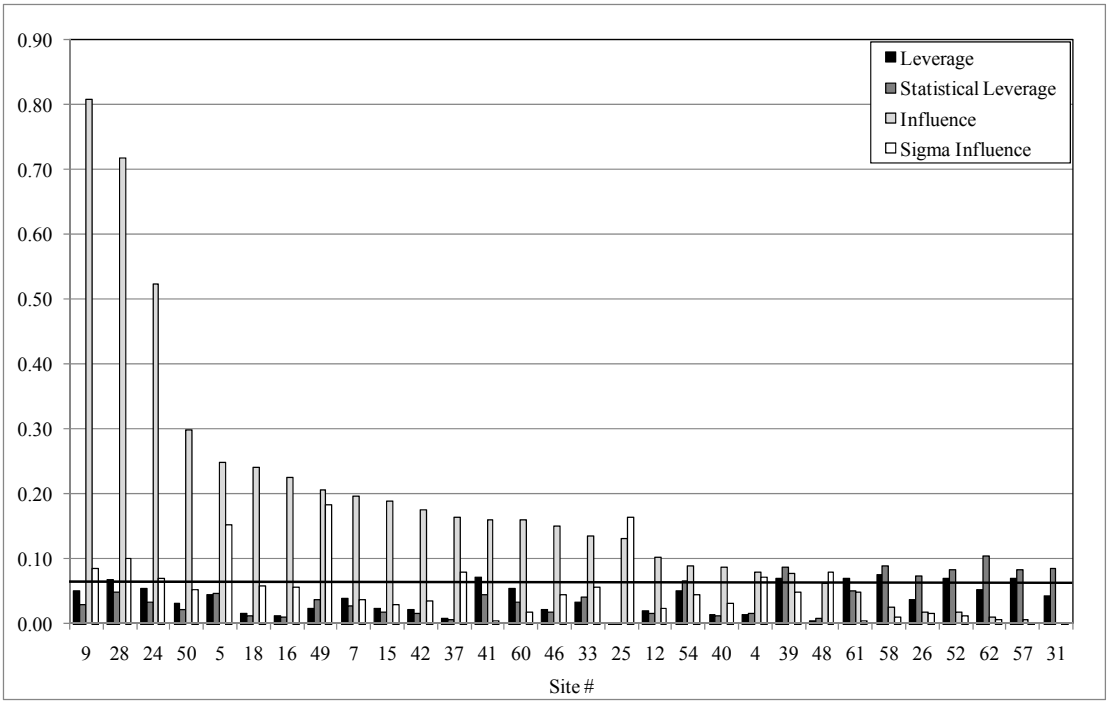

Figure 3.2: Regression Diagnostics: Leverage and influence for Illinois River Basin data set B-GLS 2 Model. The solid, horizontal line represents the threshold for high leverage and high statistical leverage, as well as, for high influence and high σ-influence.

As shown in Figure 3.1, Site 9 has extremely high influence, but only a slightly large σ-influence. Site 9 has a centered ln(Slope) value of -1.8, a record length of 86 years, a skew of -1.2, and a residual of -0.69. As shown in B-GLS 2 w/o Site 9 in Table 3.3, the resulting model error variance is 0.12, a slight decrease in the model error variance from B-GLS 2. However the biggest impact can be seen in the AVP_{new}, which decreased from 0.14 in B-GLS 2 to 0.13 in B-GLS 2 w/o Site 9. This allowed for the effective record length to increase to 57 years.

The final model in Table 3.3, is B-GLS 2 w/o Site 5. Site 5 has both very high influence, as well as extremely high σ-influence. Site 5 has a centered ln(Slope) value of 0.76, a record length of 32 years, a skew of 1.0, and a residual 1.3, the fourth largest residual among the 62 sites used in the regression. With such a large residual, Site 5 is

bound to have a large impact on an analysis. Generally, sites with small leverage and shorter record lengths, like Site 5, should not have a large influence in the regression unless they have very large residuals. Indeed, Site 5 has a large residual. When Site 5 is removed, the model error variance decreases from 0.13 to 0.12. The AVP_{new} was reduced from 0.14 to 0.13. It is important to note that the new analysis indicated that a simple constant model would be appropriate because the 95% credible region of the coefficient of ln(Slope) contains zero (Bayesian p-value = 7.3%). For comparison, the regional constant model without covariates was fitted with the entire data set B-GLS Constant (62 sites) and appears in Table 3.3.

The pseudo ANOVA table for Illinois, Table 3.4, shows that the sampling error is two to three times greater than the model error, EVR > 2. Our Illinois analysis does not explain as much of the variation in the data as the South Carolina analysis. This is evident in the R_δ^2 statistics presented in the bottom of Tables 3.2 and 3.4. The R_δ^2 value for the B-GLS 2 model for Illinois is only 0.12, while the R_δ^2 value for the B-GLS 2 model for South Carolina is 0.30.

Feaster and Tasker [2002] performed a regional analysis of the skew coefficient for the State of South Carolina. They first employed an OLS regression analysis using 102 stations and concluded the relationship between the station skews and the basin characteristics, represented by the traditional R^2 statistic, was too weak. They then computed a record-length weighted skew for two groups of sites, using the same 89 sites employed here. The reported regional skew for the Piedmont region was equal to –0.19 with mean square error of 0.090. For the other three regions, the regional skew was 0.082 with mean square error of 0.11. Tables 3.1 and 3.2 show that a statistically more complete GLS regression analysis provides a different and more statistically valid interpretation of the South Carolina regional skew data.

3.4 Conclusion

The quasi-analytic Bayesian analysis of a GLS regression model described by Reis et al. [2005] has been developed into an operational GLS regional hydrologic regression methodology. Regression diagnostic statistics for WLS and GLS models include pseudo Analysis of Variance tables, a pseudo adjusted R_{δ}^{2}, Error Variance Ratio (EVR) and Misrepresentation of the Beta Variance (MBV), leverage and influence, and σ-influence.

Two examples of regionalization of the log-space skew, the shape parameter of the Log-Pearson Type III distribution, illustrate use of the methodology. Results obtained from OLS, WLS, and GLS analyses were compared, as well as the results using the Bayesian and method of moments model-error-variance estimators. The OLS analysis provides misleading results because it does not make a distinction between the variance due to the model error and the variance due to time sampling error. The GLS analysis was the best framework because the cross-correlation of the skews, which is neglected by a WLS analysis, proved to be important. Both of these examples demonstrate that the true model error variance for regional skew models is on the order of 0.15 or less. Leverage, influence and σ-influence statistics were very useful in identifying stations that actually did have a significant impact on the analysis.

REFERENCES

Feaster, T.D. and G.D. (2002), Tasker, Techniques for Estimating the Magnitude and Frequency of Floods in Rural Basins of South Carolina, 1999, Water Resources Investigations Report 02-4140, U.S. Geological Survey: Columbia, South Carolina.

Griffis, V. W. , J.R. Stedinger, and T. A. Cohn, (2004), LP3 Quantile Estimators with Regional Skew Information and Low Outlier Adjustments, Water Resources Research, 40, W07503, doi:1029/2003WR002697.

Hardison, C.H., (1975), Generalized skew coefficients of annual floods in the United States and their application, *Water Resour Res*., 11(6), 851-854.

Interagency Advisory Committee on Water Data, (1982), Guidelines for Determining Flood Flow Frequency, Bulletin #17B, U.S. Department of the Interior, U.S. Geological Survey, Office of Water Data Coordination, Reston Virginia.

Martins E. S., and J. R. (2002), Stedinger, Cross correlations among estimators of shape, Water Resour. Res., 38 (11), 1252, doi:10.1029/2002WR001589.

McCuen, R.H., (1979), Map Skew???, *J. Water Resour. Plan and Manage. Div.*, ASCE, 105(WR2), 265-277 [with Closure 107(WR2), 582, 1981].

McCuen, R.H., (2001), Generalized flood skew: map versus watershed skew, *J. Hydrologic Eng.*, ASCE, *Vol*. 6(4), 293-299.

McCuen, R. H. and Smith, E., (2008), Origin of Flood Skew, Journal of Hydrologic Engineering, 12(9), doi: 10.1061/(ASCE)1084-0699(2008)13:9(771).

Reis, D. S., Jr., J. R. Stedinger, and E. S. Martins, (2005), Bayesian generalized least squares regression with application to log Pearson type 3 regional skew estimation, *Water Resour. Res.*, *41*, W10419, doi:10.1029/2004WR003445.

Stedinger, J.R. and G. Tasker, (1986b), Regional hydrologic analysis, 2: Model-error estimators, estimation of sigma and log-Pearson Type 3 distributions, *Water Res. Research*, 22(10), 1487-1499, 1986.

Stedinger, J.R., R.M. Vogel, and E. Foufoula-Georgiou, (1993), Frequency Analysis of Extreme Events, in Handbook of Hydrology, chao. 18, pp. 18.2 - 18.66, McGraw-Hill, New York.

Tasker, G.D., (1978), Flood frequency analysis with a generalized skew coefficient, *Water Resources Research*, 14(2), 373-376.

Tasker, G.D., and J.R. Stedinger, (1986), Regional skew with weighted least squares regression, *Journal of Water Resources Planning and Management,* 112(2), 225-237.

Tasker, G.D., and J.R. Stedinger, (1989), An Operational GLS Model for Hydrologic Regression, *Journal of Hydrology, 111*(1-4), 361-375.

CHAPTER 4

ANALYSIS OF FLOOD DATA TO SUPPORT REGIONAL FLOOD ESTIAMTES FOR THE SOUTHEASTERN UNITED STATES

4.1 Introduction to the Southeastern U.S. Study

This skew study was done in conjunction with a flood frequency effort carried out by the USGS Water Science Centers for the states of Georgia, North Carolina and South Carolina. Those states updated their estimates of flood magnitude and frequency for rural ungauged basins in the region using multi-state data [Gotvald *et al*. 2009, Feaster *et al*, 2009 and Weaver *et al*. 2009]. This multi-state approach allowed the states to regionalize over a broad area instead of state boundary lines; thus enabling a more continuous study for the Southeastern U.S. region.

As explained in Chapter 1, generalized skew coefficients are an integral piece of the US *Bulletin 17B* procedure for flood frequency analysis. Thus, the Bayesian Generalized Least Squares (B-GLS) regional skew regression developed here served to help update their flood frequency estimates.

4.2 Summary of the Southeastern U.S. Data

4.2.1 Gauge Stations

This study is based upon annual peak flow data from 489 stream flow gauges (sites) spread across seven states in the Southeastern United States. They were recommended by the United States Geological Survey's (USGS) Water Science Centers responsible for three states, Georgia, North Carolina, and South Carolina. Figure 4.1 displays the gauge sites on a map of the Southeastern US. Table 4.1 shows the breakdown of sites in each state.

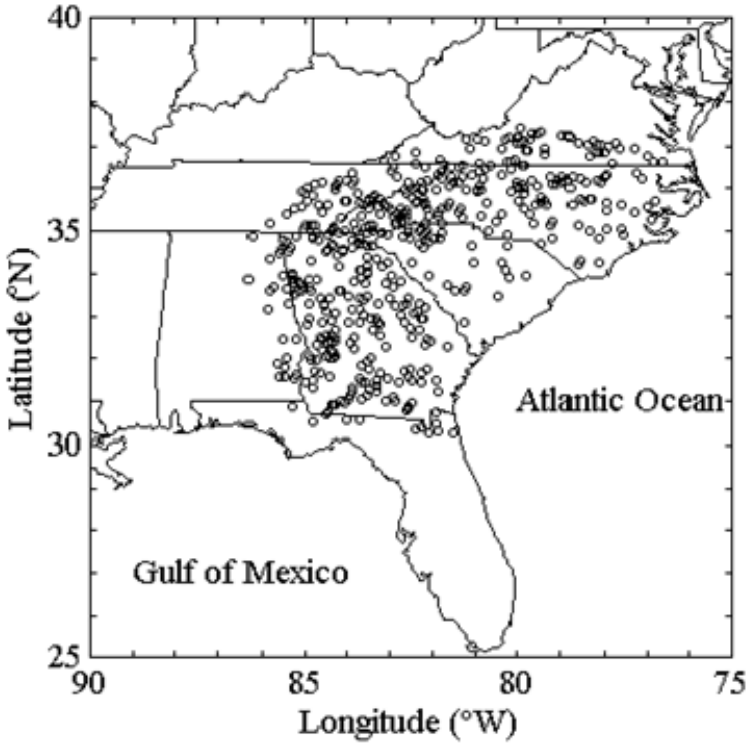

Figure 4.1: Map of the 489 gauge sites used in Southeastern U.S. study

Table 4.1: Gauge sites in Southeastern U.S. study broken-down by state

State	Number of Sites
Alabama	25
Florida	19
Georgia	169
North Carolina	127
South Carolina	38
Tennessee	64
Virginia	47

The annual peak flow data was downloaded from the USGS National Water Information System: Web Interface (NWISWeb). Each site is cataloged by a unique USGS eight or nine digit Hydrologic Unit Code (HUC), which is referred to in this study as a 'USGS site number' or simply just a 'site number'. In addition to the

USGS site number, each site was also assigned a unique index number for this study ranging from 1 to 489. A list of the 489 sites can be found in Appendix A.

4.2.2 Basin Characteristics

In addition to the peak flow data, basin characteristics for the 489 sites were provided by the USGS Water Science Centers. Table 4.2 lists the available basin characteristics.

Table 4.2: Basin characteristics for Southeastern U.S. study

General Category	Basin Chacteristics
Location of Basin Centroid:	•Latitude of Centroid (decimal degrees) •Longitude of Centroid (decimal degrees)
Basin Area:	•Drainage Area, DA (square miles) •Basin Shape Factor [Basin Length^2/DA] (unitless)
Basin Length:	•Main Channel Length (miles) •Basin Perimeter Length (miles)
Basin Slope:	•Main Channel Slope (feet/mile) •Average Basin Slope (%)
Basin Elevation:	•Average Basin Elevation (feet) •Maximum Basin Elevation (feet)
Basin Precipitation:	•Average Annual Precipitation (inches)
Basin Coverage:	•Impervious Surface Coverage (%) •Forest Coverage (%)
Basin Soil Characteristics:	•Average Soil Drainage Index for Basin (ranging from 1 to 7, with 1 denoting excessively-drained soils) •Average Hydrologic Soil Index for Basin (ranging from 1 to 4, with 1 denoting high-infiltration rates)
Physiographic Provinces within Basin:	•Blue Ridge, BR •Central Appalachians, CA •Middle Atlantic Coastal Plain, MAC •Piedmont, P •Ridge and Valley, RV •Sand Hills, SH •Southeastern Plains, SP •Southern Coastal Plain, SCP •Southwestern Appalachians, SA

The basin characteristics provided in Table 4.2 include percent of basin contained within physiographic provinces, as well as the more standard characteristics such as location of basin centroid, drainage area, main channel slope, and basin elevation.

4.3 Redundant Site Analysis

4.3.1 Introduction

In the Southeastern U.S regional skew study, it was discovered that many gauge records were for watersheds largely contained within another larger watershed represented by a different gauge. This chapter considers the impact of such nested watersheds and develops criteria for identifying redundant gauge sites, as well as developing a model for estimating the cross-correlations of the true at-site skews.

4.3.2 Cross-Correlation and Fisher Z Transformation

In order to perform a GLS analysis, an estimation of the cross-correlation of the estimator *g* of the true skew is required. As explained in Section 4.1, Martins and Stedinger [2002] used Monte Carlo experiments to determine a relationship between the cross-correlation of the skewness coefficient estimators at two sites as a function of the cross-correlation of concurrent annual maximum flows ρ. Thus, the sample cross-site correlation coefficient *r* is computed between pairs of sites using their respective concurrent annual peak flow records.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{j=1}^{n}(y_i - \bar{y})^2}} = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}} \tag{4.1}$$

Equation (4.1) is used to calculate the sample correlation r, where the x_i values are the logarithms of the annual peak flows for the first site, the y_i values are the logarithms of the annual peak flows for the second site, and n is the number of concurrent years of record between the two sites. The sample cross-correlation is restricted to the interval [-1, +1].

In order to employ a linear model with normal errors to describe the sample cross-correlation values, the cross-correlation needs to be mapped into the whole real space [-∞, +∞] to match the use of an unbounded normal-error model. The classical Fisher Z-Transformation is used to transform the sample correlation that is restricted to [-1, +1] to [-∞, +∞] so it is more amendable to being described by a regression model with an additive normal error. Because the Fisher transformation spreads out cross-correlation values near +1, one can also distinguish differences more easily. Moreover, the sampling error for the Fisher Z statistic is approximately independent of the true cross-correlation.

The Fisher Z statistic recommended by Kendall and Stuart (1961), is

$$\text{Fisher Z} = \frac{1}{2}\ln\left[\frac{1+r}{1-r}\right] \tag{4.2}$$

The variance of the Fisher Z statistic is approximately

$$Var\{\text{Fisher Z}\} = \frac{1}{n-3}, \quad \text{where } n > 50 \tag{4.3}$$

Thus, regressing on the Fisher Z statistic is easier than trying to regress on the sample correlation r for which the sampling variance is approximately (Kendall and Stewart, 1961; also Kenney and Keeping 1951, pp. 217-221)

$$Var(r) = \left(1-\rho^2\right)^2 \Big/ (n-3) \tag{4.4}$$

That the sampling variance using the Fisher Z Transformation is now independent of ρ follows from

$$\frac{dZ}{dr} = \frac{1}{2}\frac{d\left\{\ln\left[\left(1+r\right)/\left(1-r\right)\right]\right\}}{dr} = \frac{1}{2}\left[\frac{1}{\left(1+r\right)} - \frac{1}{\left(1-r\right)}\right] = \frac{1}{\left(1-r^2\right)} \qquad (4.5)$$

so that, to the first order,

$$Var(\mathrm{Z}) = \left(\frac{dZ}{dr}\right)^2 Var(r) = \frac{1}{\left(n-3\right)} \qquad (4.6)$$

4.3.3 Time Sampling Errors and Cross-Correlation of Peaks

The GLS statistical analysis depends on the estimated cross-correlations of the peak flows at different pairs of sites. These are generally estimated as a function of distances between the gauges, or in our case the distances between the centroids of the basins.

If basins are nested so one is contained within the other, then it is reasonable to expect that the cross-correlations between concurrent flood peaks would be larger than if the basins did not overlap. In theory, a cross-correlation model could be developed that captured this effect by using the observed cross-correlations between flood peaks from many basins where some of those basins are nested, and the overlap is described by new explanatory variables that are included in the model. The effect of any overlap on the cross-correlation would certainly be large when the smaller basin represents 50% or more of the larger basin. However, information on the percentage overlap, or the orientation of their overlap, or even which drainage areas are for sure contained in another is not available. Due to this lack of information regarding basin orientation,

the modeling task is much easier if it is reasonable to assume that seldom will two sites really represent mostly the same drainage area.

Thus, in order to make this assumption, a metric to identify redundant site-pairs before completing the cross-correlation model development or the Bayesian-GLS regression was needed. Section 4.3.7 describes development of that metric and its implementation.

4.3.4 Spatial Model Errors

As described in section 4.3.1, GLS regression has two errors, the time-sampling error η because of finite length records, and the model error δ intrinsic to the regression model's lack of perfection. If the cross-correlation among the concurrent floods could be model correctly, the statistical analysis would capture the cross-correlation among the time-sampling errors η.

A more troubling concern is the correlation among the model errors δ that must occur if two sites in the model represent nearly the same hydrologic experience, i.e. the two sites physically overlap, and thus, are not independent experiences. For example, this could occur if the ratios of the drainage areas DA_i/DA_j (where $DA_i > DA_j$) is equal to 1.2 when the basins are one within the other and differ by only 20% in drainage area. Then, instead of being two independent spatial observations depicting how drainage basin characteristics are related to skew (or flood quantiles), these two basins are the instead the same spatial experience (i.e. they are essentially the same basin). In that case, the statistical analysis incorrectly represents the information in the data. In the GLS regression model, each individual equation for each individual site

$$\hat{y}_i = \mathbf{x}_i\boldsymbol{\beta} + \delta_i + \eta_i \tag{4.7}$$

is intended to represent a different and unique spatial experience. It is this belief that justifies assuming that the individual δ_i are independently distributed so that the covariance matrix

$$\mathbf{\Sigma}(\delta) = \sigma_\delta^2 \mathbf{I} \tag{4.8}$$

For this assumption to be valid, it is critical that an attempt is made to retain only sites that are different spatial hydrologic experiences. If drainage areas overlap to a large extent, then this assumption is violated and the basins are no longer independent spatial experiences.

It would be possible to try to model the cross-correlation among the δ_i if basins had large overlaps, but this is difficult because δ_i is never observed. Rather, only the total errors, ε_i equal to $\delta_i + \eta_i$, are observed. Given the variance of the η_i, the variance of the δ_i can be inferred. However, the specific δ_i are not observed. [Thomas Kjeldsen and David Jones at the Institute of Hydrology (Personal communication, May 2007) have tried to infer the cross-correlation among the δ_i, but it is very difficult to resolve given the lack of precision with which the cross-correlations of the η_i are known.] Moreover, no recognized data is available delineating the degree of overlap among different basins.

In order to avoid the problems discussed in this Section and Section 4.3.3 concerning both the time sampling errors and the spatial model errors, it seems prudent to omit sites which are redundant (i.e. provide the same hydrologic experience) as characterized by a large overlap in drainage areas. This should result in no significant loss of information if the two sites have the same periods of record, because they are in fact providing the same spatial-temporal experience. If the two periods of record are not coincident, then record augmentation or record extension could be used to generate a composite site that reflects the entire hydrologic record for

this drainage basin [Maidment, pp. 17.47-17.48, 1993; Stedinger *et al*., pp. 18.36-18.37, 1993]. This new composite site should then represent a new site *k* whose fundamental spatial hydrologic experience, expressed as

$$y_k = \mathbf{x}_k \boldsymbol{\beta} + \delta_k \tag{4.9}$$

is independent of other experiences.

When omitting sites, an effort should be made to retain sites with small drainage areas and long records. We have a preference for small, unregulated sites, because that is where our regional model is most likely to be employed. However, we also have a preference for long record sites due to the greater precision with which the at-site skew can be estimated. Section 4.3.8 provides more information about the algorithm for omitting sites.

4.3.5 Screening Procedures using Normalized Distance and Drainage Area Ratios

Sections 4.3.3 and 4.3.4 motivated the importance of the identification of redundant sites. Before a decision can be made regarding how to handle two redundant sites in a regional analysis, the redundant sites must first be identified. In order to determine if two basins represent redundant sites, and thus, represent the same hydrologic experience for the purposes of conducting a regional hydrologic regression on flood quantiles or skews, two pieces of information are considered: (i) whether their watersheds are nested, and (ii) the ratio of the basin drainage areas. Thus, two screening metrics to address whether two sites are likely to be redundant are proposed. The first metric, normalized distance (ND), is used to determine the likelihood the basins are nested. The second metric, the drainage area ratio (DAR), is used to determine if two nested basins are sufficiently similar in size to conclude that they are essentially, or are at least in large part, the same watershed for the purposes of

developing a regional hydrologic model. These two metrics are explored in more detail below.

Normalized distance (ND) is a measure of the normalized, or unit-less distance, between the centroids of two basins. This metric was created in an effort to use the physical characteristics of two basins (i.e. the size of their drainage areas and distance between their centroids) to determine if the two basins are likely to be nested. ND is defined as

$$ND = \frac{D_{ij}}{\sqrt[4]{DA_i * DA_j}} = \frac{D_{ij}}{\sqrt{\sqrt{DA_i * DA_j}}} \tag{4.10}$$

where D_{ij} is the distance between the centroids of basin *i* and basin *j*.

Normalized distance is used to evaluate if two sites are likely to be nested, and thus are in large part physically redundant. It does not recognize that a small basin is located inside a much larger basin thus experiencing critically different hydrologic events.

In conjunction with normalized distance, drainage area ratios DAR are employed to determine if the size of two basins, when one basin is contained in the other, is sufficiently different that the events that generate the annual maximum floods in each basin are likely to be different. The drainage area ratio is:

$$DAR = \text{Max}\left[\frac{DA_i}{DA_j}, \frac{DA_j}{DA_i}\right] \tag{4.11}$$

If the drainage areas are different by more than a factor of *f* (i.e. DAR > *f*, where *f* = 5 was adopted in the Southeastern U.S. study described in the next section), then the fact that one basin is inside the other may not matter because they are of

sufficiently different in size that the critical characteristics of the basins and of hydrologic events that generate the annual maximum flood are likely to be different. Or, at least a large correlation between the annual floods at the two sites will not result, even if they are due to the same meteorological event because the distribution of rainfall over each basin area and the transformation of that rainfall into flow in each basin produce very different peak annual flood events. However, cross-correlation among concurrent annual peaks is not the issue here because it is dealt with explicitly in the GLS analysis. Redundancy addresses the case that the basins reflect in large part the same watershed and watershed-meteorological coupling. The correct threshold for DAR is an important issue to investigate, and could depend upon the size and characteristics of storms and the topography of the region. Section 4.3.5.1 considers an analysis of the Southeastern U.S. data set that suggests that DAR > 5 is a good threshold.

The idea of using ND and DAR to screen for redundant sites is motivated here through several simple examples. Imagine that the drainage area of Basin *A*, denoted DA_A, can be thought of as a rectangle with dimensions 1 by *r*, where $r \geq 1$. Similarly, the drainage area of Basin *B*, denoted DA_B, can be thought of as a rectangle with dimensions *w* by *w*r,* where $w \geq 1$ so that the ratio of the sides is still 1:*r*. Figure 4.2 depicts these two basins as disjoint, independent basins.

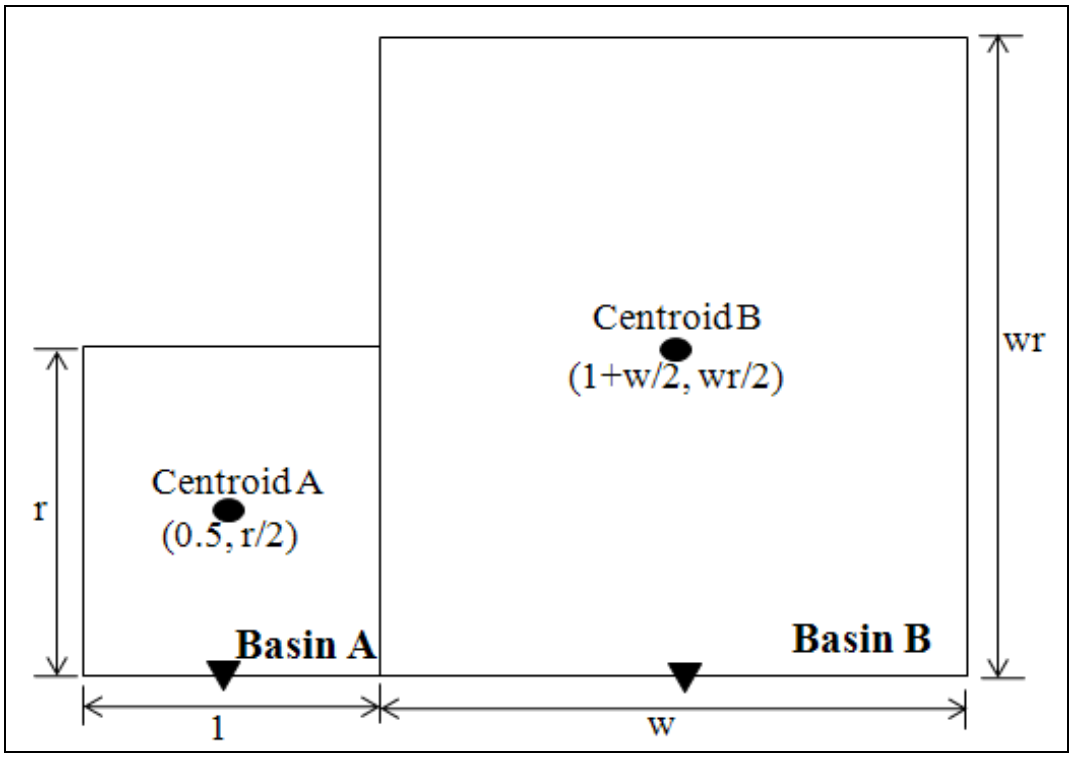

Figure 4.2: Depiction of two hypothetical distinct basins. The basin outlets are indicated by the black triangles, while the basin centroids are indicated by the black ovals.

By changing the values of *r* and *w,* the drainage areas and the length-to-width ratios of the two basins can be varied reflecting different basin configurations and geometry. The ND and DAR as a function of *r* and *w* follow from the dimensions and centroids depicted in the figure are given by

$$\begin{aligned} ND_{Distinct} &= \frac{D_{A,B}}{\sqrt{\sqrt{DA_A * DA_B}}} \\ &= \frac{\left[\left(0.5-(1+w/2)\right)^2+\left(r/2-wr/2\right)^2\right]^{1/2}}{\sqrt{\sqrt{\left(r*w^2r\right)}}} = \frac{\left[(w+1)^2+r^2(w-1)^2\right]^{1/2}}{2\sqrt{wr}} \end{aligned} \tag{4.12}$$

$$DAR = \frac{DA_B}{DA_A} = \frac{w^2 r}{r} = w^2 \tag{4.13}$$

If two basins of equal area ($w = 1$) and thus with equal dimensions are considered, then as the length-to-width ratio r increases, the normalized distance $\mathrm{ND} = 1/\sqrt{r}$, approaches zero. On the other hand, if the length-to-width ratio r is fixed, Figure 4.3 shows that as the relative size w of the large basin increases, the normalized distance ND steadily increases. Moreover, the larger the length-to-width ratio r, the faster the increase in the value of the normalized distance as w increases. This occurs because for large r, the distance between the two distinct basins looks like $(r * w)/2$ for large r and large w, so that $\mathrm{ND} = \sqrt{(r * w)}/2$; whereas for $r = 1$ corresponding to two squares, $\mathrm{ND} = \sqrt{(1 + w^2)/(2w)}$ looks like $\mathrm{ND} = \sqrt{w/2}$ for large w.

In most cases in Figure 4.3, ND > 0.5. The exceptions occur with $r > 4$ and w near 1; with long skinny basins of nearly equal size, arbitrarily small ND values are obtained.

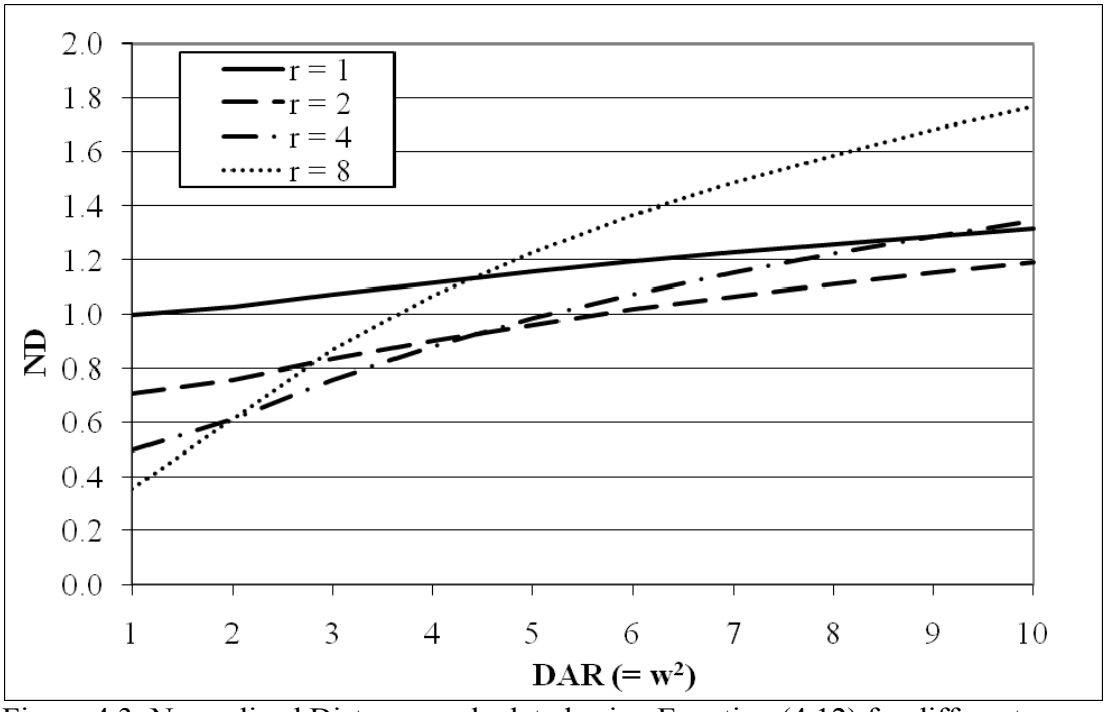

Figure 4.3: Normalized Distances calculated using Equation (4.12) for different combinations of r and w when two basins are distinct as shown in Figure 4.2.

Alternatively, consider two nested basins, where Basin *A* is smaller and wholly contained within Basin *B* (See Figure 4.4). This is different from the scenario proposed in Figure 4.2, in which Basin *A* and Basin *B* were distinct basins.

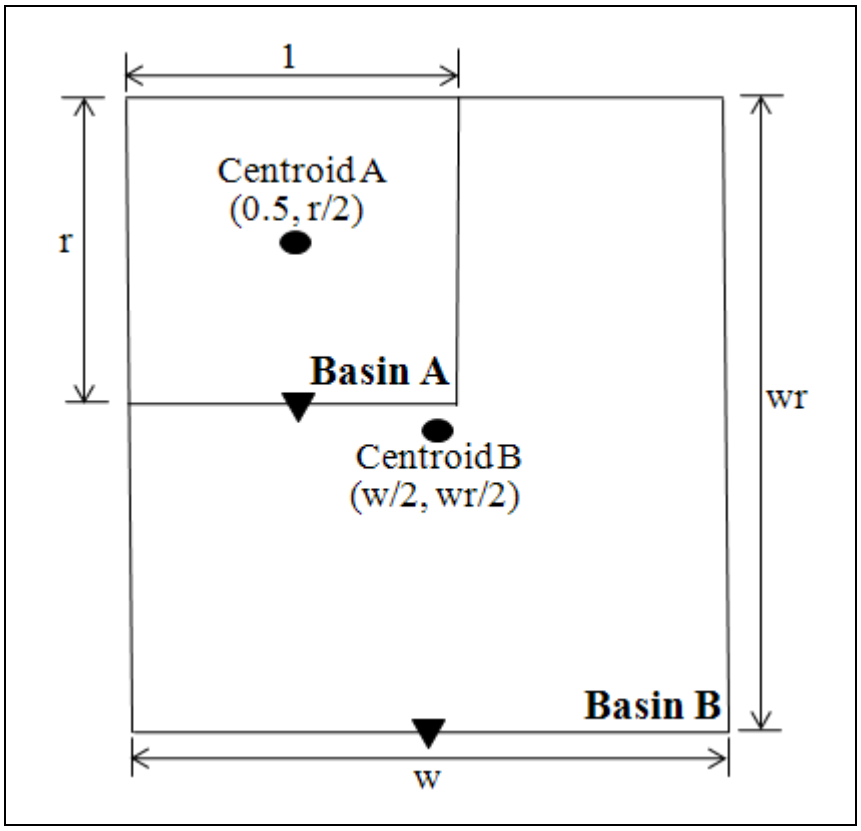

Figure 4.4: Depiction of two hypothetical, redundant basins (Basin A flows into Basin B). The basin outlets are indicated by the black triangles, while the basin centroids are indicated by the black ovals.

The geometry proposed in Figure 4.4 is a worst-case scenario in terms of maximization of ND for two nested basins. If instead of being in the corner, Basin *A* was located closer to the center of Basin *B*, it would more readily be detected as nested because the ND would be smaller because their centroids would be closer. However, by imagining Basin *A* to be nested and located as far from the centroid of Basin *B* as possible, it allows for a larger ND, and thus illustrates the difficulty of identifying nested basins with ND. On the other hand, this doesn't have to be viewed as a

weakness, but instead and interpretation of what defines redundant basins. If the centroids of two nested basins have a large ND and a large DAR, it could simply mean that the basins, although nested, are far enough apart and of sufficiently different sizes that they represent two different physical hydrologic experiences, and thus are not redundant.

The normalized distances ND for the nested basins shown in Figure 4.4 can be calculated from the dimensions and centroids that appear there:

$$\begin{aligned} \mathrm{ND}_{\mathrm{Nested}} &= \frac{\mathrm{D}_{A,B}}{\sqrt{\sqrt{\mathrm{DA}_A * \mathrm{DA}_B}}} \\ &= \frac{\left[\left(0.5 - w/2\right)^2 + \left(r/2 - wr/2\right)^2\right]^{1/2}}{\sqrt{\sqrt{\left(r * w^2 r\right)}}} = \frac{\left[\left(w-1\right)^2\left(1+r^2\right)\right]^{1/2}}{2\sqrt{wr}} \end{aligned} \quad (4.14)$$

The formula for the DAR is the same. As shown in Figure 4.4, if $w = 1$ then $\mathrm{DA}_A = \mathrm{DA}_B = \mathrm{r}$, so that Basin *A* and Basin *B* are in fact the same basin, and thus their ND = 0. Thus, the smallest normalized distances are obtained when the nested basins are the same size, as shown in Figure 4.5. As *w* increases, ND increases. And for any $w > 0$, ND also increases with *r*. Again, smaller ND values are obtained if Basin *B* were located closer to the center of the larger Basin *A*.

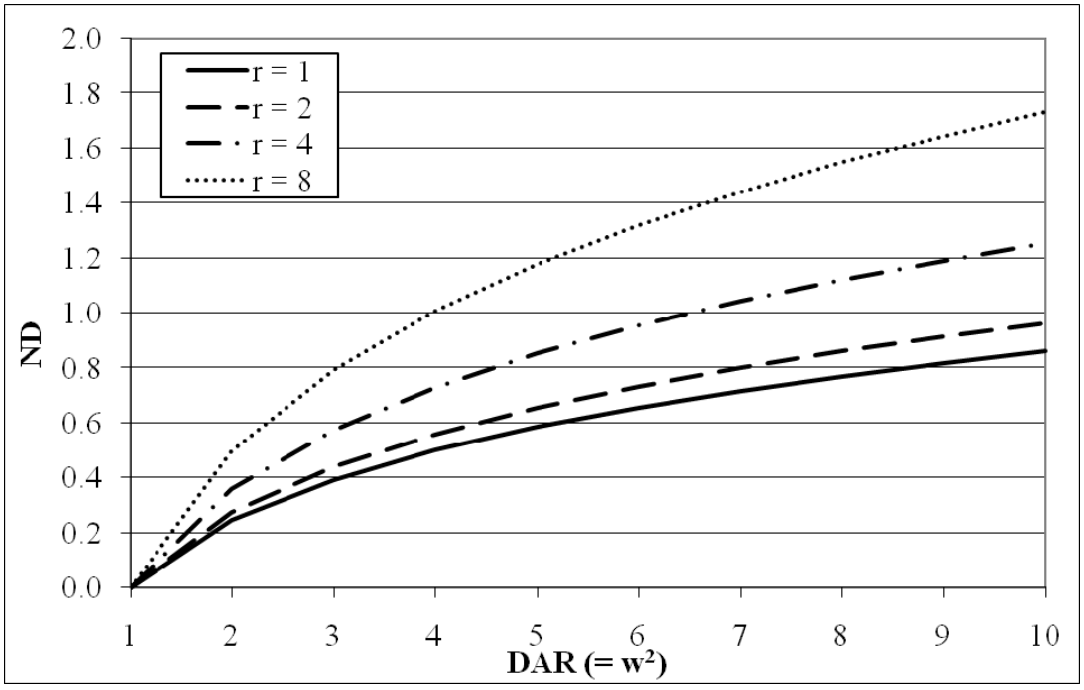

Figure 4.5: Normalized Distances for different combinations of *r* and *w* when basins are nested as shown in Figure 4.4.

By comparing Figures 4.3 and Figure 4.5, a better understanding of the behavior of normalized distance metric can be obtained as it relates to the relative position of two basins. For example, when two basin are distinct neighbors, as depicted by Figure 4.2, if the length to width ratio, *r* is 4:1 or more, the normalized distance drops to 0.5 or less and by that threshold would be incorrectly classified as redundant. Also, Figure 4.3 shows two distinct basins can have a ND < 0.5, when $r \geq 8$ and DAR ≤ 1.5. It is in this instance that if the threshold for classifying pairs of sites as redundant is set at ND < 0.5 and DAR < 5, that distinct basins would be incorrectly classified as redundant.

On the other hand, when one basin is nested within another according to the worst-case geometry depicted in Figure 4.4, if $w = 1$ for any *r*, they are in fact the same basin and thus ND = 0. As shown in Figure 4.5, when one basin is contained within the other there are many more possible ways that ND < 0.5 could result. For example, ND ≤ 0.5 when

- $r = 1$ and $w \leq \sqrt{4} = 2$;
- $r = 2$ and $w \leq \sqrt{3.5} = 1.9$;
- $r = 4$ and $w \leq \sqrt{2.5} = 1.6$;
- $r = 8$ and $w \leq \sqrt{2} = 1.4$;

Thus, if ND is less than 0.5, it is clearly possible that the two drainage areas are nested with the larger basin containing the smaller basin of they are of similar size. If basins are tall and narrow (i.e. $r \geq 8$), or very different in size, then it is possible for them to be distinct even though their ND ≤ 0.5. Fortunately, this is a worst case configuration for nested sites. If the smaller site is near the center of the large site, ND can be very much smaller, or even zero.

Real basins come in a range of sizes and shapes. Section 4.3.6 provides examples of real basins. Although there are times when the combination of normalized distance and drainage area ratio will incorrectly identify distinct basins as redundant, it appears that ND < 0.5 can successfully be used as a screening metric to identify redundant basins which represent the same hydrologic experience. Examination of detailed maps which document watershed boundaries can be used to confirm whether or not basins are in fact nested. This step can illuminate basins that are distinct, but which might have been classified as redundant. This should eliminate false positives. The error that might be made is to fail to recognize redundant pairs because of a large *r* with a modest DAR resulted in a ND < 0.5. Because large *r* can results in very small ND values, these false negatives could be a concern.

4.3.6 Redundant Site Example Using Data from South Carolina

The concept of redundant site and the thresholds for ND and DAR can best be understood by looking at a real example. Figure 4.6 was produced using the USGS NHD Geodatabase. It presents three peak flow gauges, shown as black circles, located

in South Carolina. The rivers and their main tributaries are represented by the thick gray lines, while the outlines of each of the three basins are represented by thick black lines. The gauge inside Basin *A* is located on the South Fork Edisto River, the gauge inside Basin *B* is located on the North Fork Edisto River, and the gauge inside Basin *C* is located on the Edisto River after the confluence of the South Fork and the North Fork.

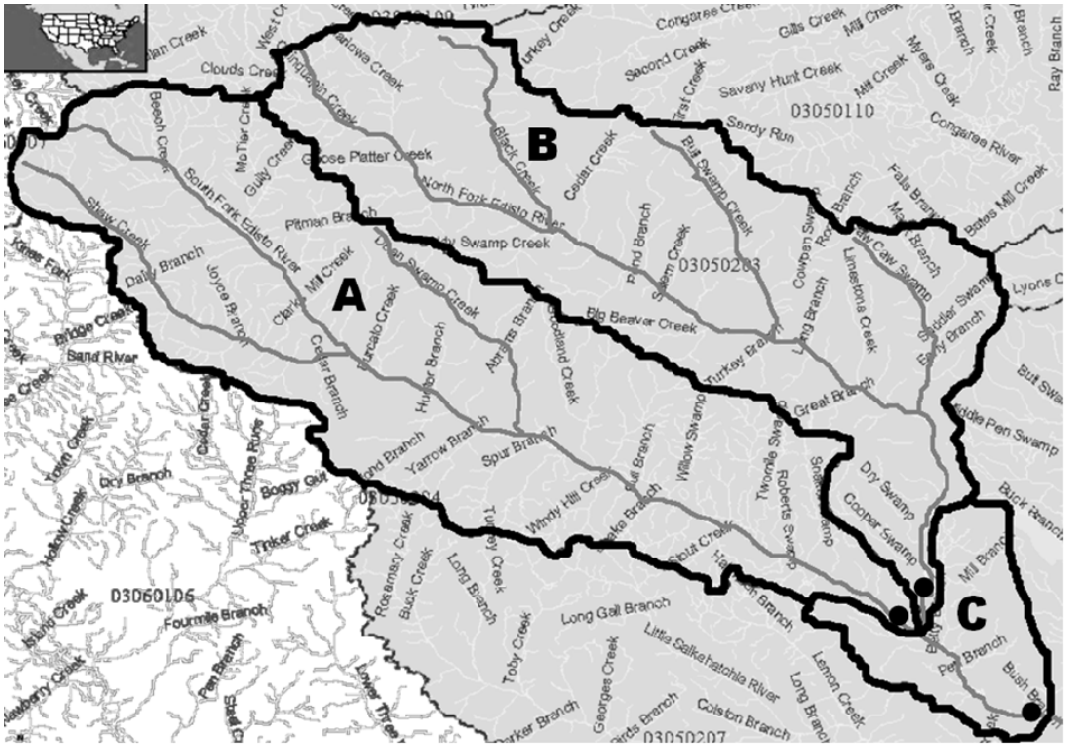

Figure 4.6: Illustration of redundant drainage basins in South Carolina on the Edisto River. The map was produced using the USGS NHD Geodatabase from http://nhdgeo.usgs.gov/viewer.htm.

It is easy to see from Figure 4.6 that the South Fork and the North Fork of the Edisto River are two separate drainage basins, denoted Basin *A* and Basin *B*, respectively. The drainage area of Basin *A* is 734 mi^2 and the drainage area of Basin *B* is 686 mi^2. The main channel length of Basin *A* is roughly 67 mi and the main channel length of Basin *B* is roughly 65 mi. The average basin widths for both Basin

A and Basin *B* are roughly 16 mi. Thus, the width to length ratios of Basin *A* and Basin *B* are 1 to 4.2 and 1 to 4.1. The centroids of Basin *A* and Basin *B* are roughly 18 mi apart, thus their ND = 0.69 and their DAR = 1.0. One criterion discussed above, and employed latter in the analysis, is to assume that basins with ND >0.5 are not nested, and that is shown to be true in this case.

However, if instead the focus is placed on Basin *C*, it is apparent that the three basins overlap, or more specifically Basins *A* and *B* are contained within Basin *C*. This is due to the fact that Basin *C* contains the confluence of Basins *A* and *B*. Basin *C* has the following basin characteristics: drainage area = 1726 mi^2, channel length = 109 mi, average channel width 43 mi, and basin width to length ratio = 1 to 2.5.

If Basin *A* and Basin *C* are compared, it is evident that Basin *A* is about 43% of the drainage area of Basin *C* and their centroids are only 12 mi apart. They have a ND = 0.36 and a DAR = 2.4. Thus, using the screening methods suggest above, these two sites would be classified as redundant because ND < 0.5 and DAR <5. As shown in Figure 4.6, Basin *A* is in fact contained in Basin *C*, and thus these two basins should not be viewed as independent hydrologic experiences. The same conclusion is drawn when comparing Basin *B* and Basin *C*. Basin *B* is about 40% of the drainage area of Basin *C* and their centroids are only 9 mi apart. Thus, they have a ND = 0.26 and a DAR = 2.5, both of which are below the thresholds indicating that these two basins are redundant. As will be shown later, in the case of the three basins above, Basin *C* was removed from the analysis in order to address the issue of redundant sites, while Basin *A* and Basin *B* were retained.

4.3.7 Screening Procedure and Results

By using normalized distance ND and the drainage area ratio DAR, described above in Section 4.3.5, a screening procedure for redundant sites can be formulated.

In order to develop the appropriate thresholds for both ND and DAR for the Southeastern U.S. data set using the cross-correlations of the concurrent peak annual flows, only those sites whose concurrent record lengths were greater than 30 years were considered. After thresholds were chosen, the screening procedure was then applied to the entire data set, including those pairs of sites with less than 30 years of concurrent records. Figure 4.7 describes the algorithm used to screen sites,

If ND< T_{ND} **&** DAR < T_{DAR}
    IF (DA_{small} has >= 30 yrs of data), **THEN** (remove DA_{small})
    ELSE IF [(DA_{small} <30 yrs of data) **&** (record length of DA_{small} + 5)] ≥
        (record length of DA_{large}), **THEN** (remove DA_{small})
ELSE (remove DA_{small})

Figure 4.7: Screening algorithm for redundant sites

where T_{ND} is the normalized distance threshold, and T_{DAR} is the drainage area ratio threshold, DA_{small} is the site with the smaller drainage area, and DA_{large} is the site with the larger drainage area. The screening algorithm in Figure 4.7 first identifies troublesome pairs: pairs of sites whose ND is less than T_{ND} and whose DAR is less than T_{DAR}. After identifying the troublesome pairs, the algorithm then runs through those troublesome pairs in index number order (see Section 4.2.1 for description of index number) to resolve each redundant site conflict by recommending the removal of one of the two sites. There are cases where 3 or even 4 sites are in conflict, and the order of the comparison can affect the number of sites removed and the selection of sites to remove. As a simple illustration if the sizes of basins B1, B2 and B3 are such that B1 < B2 < B3, then it may be sufficient to remove B2, or both B1 and B3. The algorithm used in the Southeastern U.S. study considers sites only two at-a-time.

In order to determine the appropriate ND and DAR thresholds, different combinations were tried and their results examined. Figures 4.8 and 4.9, present the

cross-correlation and the Fisher Z-transformation of the cross-correlations versus ND. The figures show the suspected redundant troublesome pairs of sites as open circles and the non-redundant pairs of sites as solid circles (or dots) for the proposed selection criteria. The figures include all site-pairs with at least 30 years of concurrent record. Many pairs of sites in this study fail to meet the ND and DAR criteria and thus, one of them was omitted.

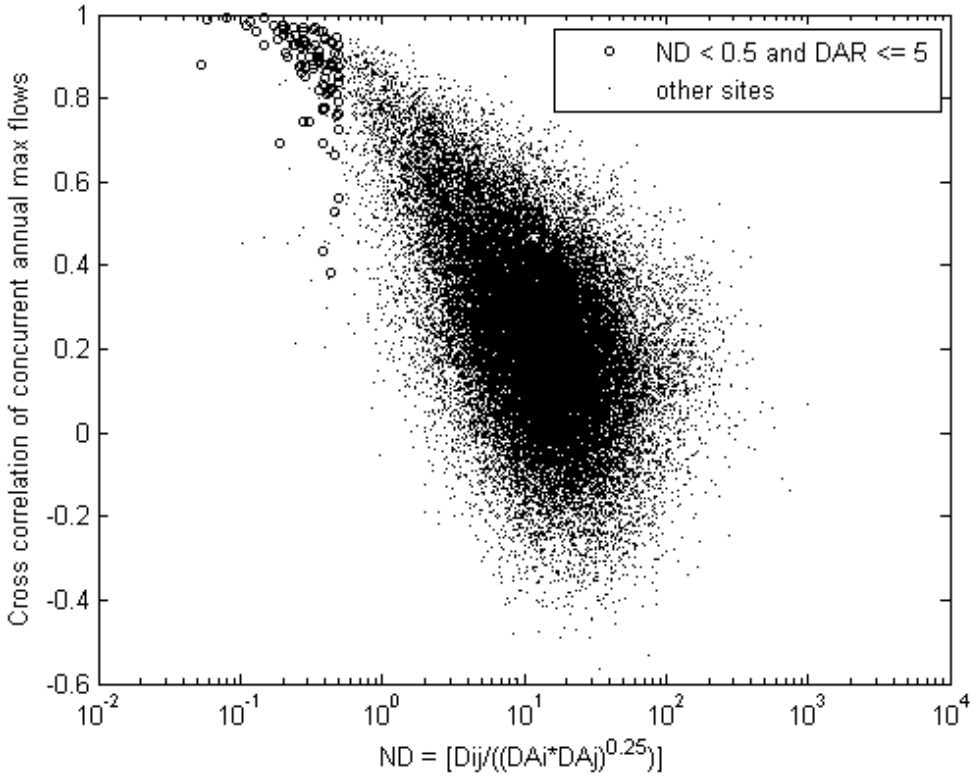

Figure 4.8: Sample cross-correlation versus normalized distance ND for site pairs with greater than 30 years of shared record. ND and drainage area ratio DAR can identify site-pairs with unusually large cross-correlation which are probably due to redundancy.

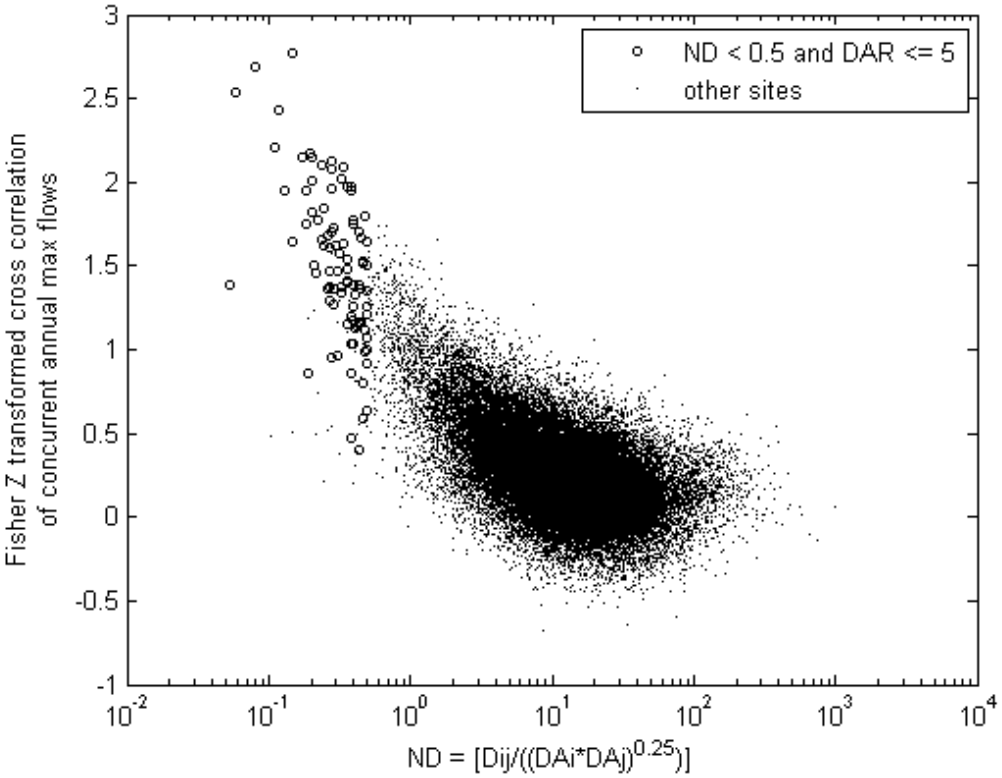

Figure 4.9: Sample Fisher Z transformed cross-correlation versus normalized distance ND for site pairs with greater than 30 years of shared record. ND and drainage area ratio DAR can identify site-pairs with unusually large cross-correlation which are probably due to redundancy.

In Figure 4.9, a Fisher Z transformed cross-correlation Z = 3 corresponds to a cross-correlation $r = 0.995$, Z = 2 corresponds to $r = 0.964$, Z = 1.5 corresponds to $r = 0.905$, and Z = 0 corresponds to $r = 0$. As demonstrated by the figures, it appears that a threshold T_{ND} equal to 0.5 for ND and a threshold T_{DAR} for DAR equal to 5 would remove most of the site-pairs with extremely high cross-correlations, and it removes all of those redundant site-pairs with cross-correlations larger than Z = 1.75 or $r = 0.941$.

The screening procedure outlined in Figure 4.7 was applied to the 489 sites comprising the Southeastern U.S. data set to identify pairs of sites whose ND < 0.5

and whose DAR < 5. For sites whose concurrent record length is greater than 30 years, the screening procedure identifies 94 troublesome pairs, thus leading to the recommendation that 61 individual sites be omitted from the study. After running the screening metric on the entire data set, 92 sites are suggested for omission from the study. (This includes the 61 sites removed previously whose concurrent record length is greater than 30 years.) Table 4.2 and Table 4.3 show the breakdown of both the omitted sites and the sites remaining in the study by both state and physiographic region. The table of troublesome pairs of sites as well as those sites omitted from the study to resolve the problem of duplicate basins can be found Appendix A.

Table 4.3: Breakdown of sites in Southeastern U.S. study by state after redundant site screening.

	AL	FL	GA	NC	SC	TN	VA	Total
# of Sites Included in Study:	18	12	142	107	28	52	38	397
# of Sites Removed from Study:								
Identification and Screening	7	7	27	20	10	12	9	92
# of Total Sites	25	19	169	127	38	64	47	489

Table 4.4: Breakdown of sites in Southeastern U.S. study by physiographic province after redundant site screening.

	BR	CA	MAC	P	RV	SH	SP	SCP	SA	Total
# of Sites Included in Study:	72	2	15	139	42	17	79	19	12	397
# of Sites Removed from Study:										
Identification and Screening	14	1	2	44	8	2	14	5	2	92
# of Total Sites	86	3	17	183	50	19	93	24	14	489

Note: Region for each site is chosen according to which region the largest % of basin is contained in.

As shown in Table 4.3, Georgia has the largest number of sites (142) followed by North Carolina (107), while Alabama (18) and Florida (12) have the fewest number of sites. One concern is that this study focuses on Georgia, North Carolina, and South

Carolina and only 38 sites from South Carolina are included in the data set, ten of which had to be removed due to redundant sites.

Table 4.4 shows the breakdown of sites by physiographic province. It is important to note that physiographic province in Table 4.4 was categorized according to the province with the largest percent of the basin. The Piedmont province (P) has by far the largest number of sites (139) in the study, while the Central Appalachian province (CA) has the fewest number of sites (2) in the study. With only 2 sites in the study with the majority of their basin area located in the Central Appalachians, it seems difficult to define the province as its own region. This is further addressed when the physiographic provinces are used as regression parameters in the Bayesian GLS regional skew regression.

4.4 Cross-Correlation Model for Regional B-GLS Skew Regression

4.4.1 Introduction

Bulletin 17B entitled "Guidelines for Determining Flood Flow Frequency", recommends the log-Pearson Type III distribution to define the frequency distribution of annual flood series. This distribution has three parameters, which are normally taken to be: the mean, the standard deviation and the skew. This thesis focuses on the third parameter, the skew. Reis *et al.* [2005] demonstrated that regional regression models can be used to develop regional skew estimators *Bulletin 17B* prescribes how such regional skewness estimators can be used in conjunction with at-site skews to produce a weighted skew estimator. In order to implement the Reis *et al.* [2005] Bayesian GLS regression framework, the correlations among skewness estimators must be described. This is a part of the GLS regression framework, in which the relationships between estimators for different sites are measured by their cross-correlations. As the true skew values are not know, these correlations can be

approximated as a function of the cross-correlation of the peak flood flows as developed in Martins and Stedinger [2002].

Using Monte Carlo experiments, Martins and Stedinger [2002] developed a relationship between the cross-correlation of the skewness coefficient estimators at two sites as a function of the cross-correlation of concurrent annual maximum flows denoted ρ_{ij}. A second factor (cf_{ij}) accounts for the sample size difference between the sites, as well as the concurrent record lengths. Their cross-correlation model is

$$\hat{\rho}\left(\hat{\gamma}_i, \hat{\gamma}_j\right) = Sign\left(\hat{\rho}_{ij}\right) cf_{ij} \left|\hat{\rho}_{ij}\right|^{\kappa}, \quad \text{where} \quad cf_{ij} = n_{ij} / \sqrt{\left(n_{ij}+n_i\right)\left(n_{ij}+n_j\right)} \tag{4.14}$$

where n_{ij} is the common record period, n_i and n_j are the extra observation periods and κ is an empirical constant between 2.8 and 3.3.

In Reis *et al.*[2005], the inter-site correlation coefficient between concurrent flows $\rho(d_{ij})$ is modeled solely as a function of the distances between two site gauges. It is important to note here that the distance used by Reis *et al*. [2005] is not the distance between basin centroids as is used in this report, but instead is the distance between gauges. The gauges for each basin are located at the outlet of the basin while the centroid is the geometric center of each basin. Thus, using the distance between basin centroids should present a better representation of the proximity and similarities of any basin pair.

The model adopted by Tasker and Stedinger [1989, eq 21] to model the cross-correlation of concurrent annual max flows is

$$\hat{\rho}_{ij} = \theta^{\left\lfloor \frac{d_{ij}}{\alpha d_{ij}+1} \right\rfloor} = \exp\left\{ \left\lfloor \frac{d_{ij}}{\alpha d_{ij}+1} \right\rfloor \ln\theta \right\} \tag{4.14a}$$

where d_{ij} is the distance between site gauges and θ and α are parameters of the model in which $0 < \theta < 1$ and $\alpha > 0$. The function has a maximum value of 1 at $d_{ij} = 0$ and approaches a minimum of $\theta^{1/\alpha}$ as d_{ij} approaches infinity. This model is difficult to understand and does not offer much flexibility at $d_{ij} = 0$.

The model is perhaps more easily understood if written as

$$\hat{\rho}_{ij} = \exp\left\{\frac{\ln\theta}{\alpha}\left[\frac{d_{ij}}{d_{ij}+1/\alpha}\right]\right\} \qquad (14.14b)$$

where $(1/\alpha)$ and $(\ln\theta/\alpha)$ can be seen as the two natural parameters corresponding to a measurements at which $\left[d_{ij}/\left(d_{ij}+1/\alpha\right)\right]$ has a value of 0.5 and $(\ln\theta/\alpha)$ is equal to the logarithm of the lower bound.

However, if the distance between basin gauges is zero, then the according to Equation 4.14a and 4.14b, the cross-correlation is always one. This occurs even when the basins are not redundant. A more flexible model is developed in the next section.

In order to develop an improved estimate of the regional skew estimator and its precision, it is important to represent the cross-correlation as accurately as possible. Thus, the sub sections of Section 4.4 focus on developing an improved model of inter-site cross-correlation using peak annual flows, as well as, the distance between the centroids of basins.

4.4.2 Cross-Correlation Modeling Procedure

In modeling the cross-correlations, the Fisher Z-transformation, described in Section 4.3.2, is used to transform the sample correlation values so that their variances are independent of their cross-correlations. Also, this transformation allows for a model with normal errors, as the feasible range of cross-correlation values is expanded

out from [-1.+1] to [-∞,+∞] . In order to find the best cross-correlation model, the Fisher Z transformation of the correlations is regressed on several covariates described below.

The first covariate is represented as:

$$X_1 = \mathrm{D}_{ij}^{c} \tag{4.15}$$

or equivalently $X_1 = \left[\mathrm{D}_{ij}^{c}\right]/c$ wherein D_{ij} is the distance between basin centroids and c is the power to which the distance is raised. The expression $\left[\mathrm{D}_{ij}^{c} - 1\right]/c$ has the advantage that it is a continuous function of c. Even at D_{ij} equal to zero the expression has a value, $X_1 = -1/c$. It is expected that c will be less than one and greater than zero. As distance between basin centroids becomes very large, it becomes more likely that the basins are not affected by the same hydrologic events, and thus it is expected that their correlation would decrease toward zero. However, as the distance between basin centroids approaches zero, it is more likely that the basin would be affected by the same hydrologic events, and thus it is expected that their correlation would increase toward 1. This implies that the regression coefficient in front of this term, β_1, will be negative.

The second covariate is represented as:

$$X_2 = \ln\left(\mathrm{DA}_i * \mathrm{DA}_j\right) \tag{4.16}$$

where DA_i and DA_j represent the drainage areas of the two basins of interest. This term represents the log geometric mean of the two basins. It is suspected there is a quicker drop off (i.e. the correlation decreases at a faster rate with increasing distance) for small basins. This implies that the regression coefficient in front of this term, β_2, will be positive.

Finally the third covariate is represented as:

$$X_3 = \left| \ln\left(\frac{\mathrm{DA}_i}{\mathrm{DA}_j} \right) \right| \tag{4.17}$$

This term is a measure of the discordance; basins that are of similar size are more likely to have high cross-correlations than basins that are of very different size, holding the distance between basins D_{ij} constant. As expected, basins of very different sizes react to storm events of very different sizes. For example, a small, localized thunderstorm could produce a flash flood in a basin with a small drainage area, while that same thunderstorm could produce only minor flows in a nearby basin with a large drainage area. It would take a much larger regional storm to produce large peak flows in the large basins, while that same large storm may not affect the smaller basin to the same degree. This implies that the regression coefficient in front of this term, β_3, will be negative.

The nonlinear equation used to model cross-correlation is:

$$Z_{ij} = \alpha + \exp\left(\beta_0 + \beta_1 \mathrm{D}_{ij}^{c} + \beta_2 \ln\left(\mathrm{DA}_i * \mathrm{DA}_j \right) + \beta_3 \left| \ln\left(\frac{\mathrm{DA}_i}{\mathrm{DA}_j} \right) \right| \right) + \varepsilon \tag{4.18}$$

This equation employs the three covariates described above to develop a relationship between cross-correlation of annual peak flows and the geometry of their watersheds. It is expected that the additional geometric mean and discordance covariates will help to better explain the variations found in the inter-site cross-correlations. The results for this nonlinear regression analysis are presented in the next section. The constant α is included in the regression to account for the result that even at the largest distances represented in this study, the cross-correlation between basins does not fall to zero.

4.4.3 Cross-Correlation Model Analysis and Results for the Southeastern U.S.

The regression model is calibrated using data for pairs of sites which have at least 70 years of concurrent record. For these pairs of sites with at least 70 years of concurrent record, the average number of concurrent years of record is 75, with a standard deviation of about 5 years and a maximum of 111 years. Thus, the sampling errors of the Fisher Z transformations are relatively constant.

In the case of the Southeastern U.S. study, 92 redundant sites are omitted to resolve the problem of redundant basins in the study, and cross-correlations are then computed using all 1,317 pairs generated with the remaining 397 sites that had at least 70 years of concurrent record. Table 4.5 presents these results after the screening procedure outlined in Section 4.3.6 is performed and the redundant sites are omitted. Table 4.6 presents the results using all 3011 pairs (i.e. no redundant sites were removed from the study) that resulted from using all 489 sites with at least 70 years of concurrent record are considered. In both cases, with and without redundant sites, the Fisher transformed cross-correlations were regressed against distance between basin centroids, as well as the geometric area and discordance (drainage area ratio) as explained in Section 4.4.2. Two functions were employed to measure the distance between basin centroids; the first measured distance between basin centroids, D_{ij} in miles, while the second measured distance between basin centroids using the unitless normalized distance metric, ND_{ij}, described in Section 4.3.5.

The model computations summarized for Table 4.5 represent results after the screening procedure outlined previously is performed and the redundant sites were omitted. Meanwhile, the model computations summarized for Table 4.6 represent results without any screening procedures, and thus include all sites including those which have been classified previously as redundant sites, recall that some redundant site pairs had very large cross-correlations exceeding $r = 0.99$. The first column of

both Tables 4.5 and 4.6 depicts the models used to carry out the cross-correlation regressions. The subsequent seven columns of the table present regression coefficients, while the last three columns present measures of model fit: the estimated model error variance σ_δ^2, Pseudo R_δ^2, and effective record length (ERL) [Gruber *et al*., 2007], respectively.

The model fit measures are developed in Chapter 2, which also discusses their theoretical motivation. However, the application here is somewhat different. Here estimated model error variance σ_δ^2 is the estimated mean square error minus the average sampling error [Hardison, 1971; Stedinger and Tasker, 1986b]. This is a reasonably unbiased estimator of the true prediction error for the model. A good model should have a small error variance, implying that the regression is fairly accurate.

In the cross-correlation regression, the error of most concern is the model error variance because the sampling error is unexplainable and represents noise in the data that complicates the analysis. The Pseudo R_δ^2 statistic is a goodness-of-fit statistic in which the unexplainable sampling error variance is separated from the total error variance, leaving behind the fraction of the variance accounted for by the model error variance [Gruber *et al*., 2007]. Pseudo R_δ^2 as employed here is defined as

$$\text{Pseudo } \mathrm{R}_\delta^2 = 1 - \frac{\hat{\sigma}_\delta^2(k)}{\hat{\sigma}_\delta^2(0)} \tag{4.19}$$

where $\hat{\sigma}_\delta^2(k)$ is the estimated model error variance with k explanatory variables and $\hat{\sigma}_\delta^2(0)$ is the estimated model error variance when no explanatory variables are present. Thus, a large Pseudo R_δ^2 indicates a good model fit as this indicates that a significant

amount the fraction of the total error variance explainable by the model is explained by the model.

Effective record length (ERL), is derived from average variance of prediction (AVP), where AVP quantitatively evaluates a regression model's ability to predict skew at a new site not used in the regression. The power of the regional skew estimate is evaluated by comparing it to an equivalent at-site estimator with the calculated ERL. Thus, the larger the ERL, the more powerful the regional skewness estimator.

Focusing on the results from Table 4.5, the best models, as classified by small σ_δ^2 and large Pseudo R_δ^2, are models B, F, G, and H. However, both G and H have regression coefficients which are not significantly different from zero. After eliminating models G and H, the focus can shift to Models F and B. In Model F, the β_4 regression coefficient is negative and small. The negative value does not agree with the previous reasoning that the log geometric mean covariate should represent a faster decrease in correlation for small basins, implying a positive regression coefficient. Thus, this leaves Model B as the model of best fit with physically reasonable coefficients. Model B has a σ_δ^2 = 0.007 corresponding to an effective record length of over 150 years and a Pseudo R_δ^2 of 83%. With this model, cross-correlations vary from 0.935 when D_{ij} is equal to zero, to 0.14 for large D_{ij}.

Table 4.5: Summary of cross-correlation regressions with n = 1317 site-pairs (redundant sites omitted). The underlined regression parameters indicate that those values are not statistically different from zero at the 10% level.

MODEL	REGRESSION PARAMETERS							σ_δ^2	Pseudo R^2	ERL
	α	β_0	β_1	β_2	β_3	β_4	β_5			
A: $Z_{ij} = \alpha + \varepsilon$ where $\alpha = \beta_0 + \exp(\beta_1)$	0.306							0.040	0.0%	28
B: $Z_{ij} = \beta_0 + \exp\left(\beta_1 + \beta_2\left(\frac{D_{ij}^{\beta_3} - 1}{\beta_3}\right)\right) + \varepsilon$		0.136	0.332	-0.069	0.609			0.007	83%	152
C: $Z_{ij} = \beta_0 + \exp\left(\beta_1 + \beta_2\left(\frac{ND_{ij}^{\beta_3} - 1}{\beta_3}\right)\right) + \varepsilon$		0.180	-0.028	-0.147	0.592			0.015	62%	69
D: $Z_{ij} = \beta_0 + \exp\left(\beta_1 + \beta_4\left(\ln\left(A_i A_j\right)\right)\right) + \varepsilon$		0.233	-3.039			0.100		0.040	0.3%	28
E: $Z_{ij} = \beta_0 + \exp\left(\beta_1 + \beta_5\left\lvert\ln\left(A_i / A_j\right)\right\rvert\right) + \varepsilon$		0.183	-1.643				-0.400	0.038	4.5%	29
F: $Z_{ij} = \beta_0 + \exp\left(\beta_1 + \beta_2\left(\frac{D_{ij}^{\beta_3} - 1}{\beta_3}\right) + \beta_4\left(\ln\left(A_i A_j\right)\right)\right) + \varepsilon$		0.116	0.792	-0.106	0.496	-0.061		0.006	85%	166
G: $Z_{ij} = \beta_0 + \exp\left(\beta_1 + \beta_2\left(\frac{D_{ij}^{\beta_3} - 1}{\beta_3}\right) + \beta_5\left\lvert\ln\left(A_i / A_j\right)\right\rvert\right) + \varepsilon$		0.131	0.406	-0.079	0.572		-0.032	0.007	83%	152
H: $Z_{ij} = \beta_0 + \exp\left(\beta_1 + \beta_2\left(\frac{D_{ij}^{\beta_3} - 1}{\beta_3}\right) + \beta_4\left(\ln\left(A_i A_j\right)\right) + \beta_5\left\lvert\ln\left(A_i / A_j\right)\right\rvert\right) + \varepsilon$		0.116	0.785	-0.104	0.501	-0.062	0.005	0.006	85%	166

Table 4.6: Summary of cross-correlation regressions with n = 3011 site-pairs (redundant sites included). The underlined regression parameters indicate that those values are not statistically different from zero at the 10% level.

MODEL	REGRESSION PARAMETERS							σ_δ^2	Pseudo R^2	ERL
	α	β_0	β_1	β_2	β_3	β_4	β_5			
A: $Z_{ij} = \alpha + \varepsilon$ where $\alpha = \beta_0 + \exp(\beta_1)$	0.318							0.057	0.0%	21
B: $Z_{ij} = \beta_0 + \exp\left(\beta_1 + \beta_2\left(\frac{D_{ij}^{\beta_3} - 1}{\beta_3}\right)\right) + \varepsilon$		0.143	0.439	-0.063	0.630			0.018	69%	60
C: $Z_{ij} = \beta_0 + \exp\left(\beta_1 + \beta_2\left(\frac{ND_{ij}^{\beta_3} - 1}{\beta_3}\right)\right) + \varepsilon$		0.111	-0.301	-0.658	0.225			0.016	72%	66
D: $Z_{ij} = \beta_0 + \exp\left(\beta_1 + \beta_4\left(\ln\left(A_i A_j\right)\right)\right) + \varepsilon$		0.273	-8.982			0.422		0.055	4%	21
E: $Z_{ij} = \beta_0 + \exp\left(\beta_1 + \beta_5\left\lvert\ln\left(A_i / A_j\right)\right\rvert\right) + \varepsilon$		0.236	-2.224				-0.197	0.056	2%	21
F: $Z_{ij} = \beta_0 + \exp\left(\beta_1 + \beta_2\left(\frac{D_{ij}^{\beta_3} - 1}{\beta_3}\right) + \beta_4\left(\ln\left(A_i A_j\right)\right)\right) + \varepsilon$		0.119	-0.753	-0.133	0.444	0.117		0.013	77%	81
G: $Z_{ij} = \beta_0 + \exp\left(\beta_1 + \beta_2\left(\frac{D_{ij}^{\beta_3} - 1}{\beta_3}\right) + \beta_5\left\lvert\ln\left(A_i / A_j\right)\right\rvert\right) + \varepsilon$		0.158	0.579	-0.052	0.693		-0.133	0.016	72%	64
H: $Z_{ij} = \beta_0 + \exp\left(\beta_1 + \beta_2\left(\frac{D_{ij}^{\beta_3} - 1}{\beta_3}\right) + \beta_4\left(\ln\left(A_i A_j\right)\right) + \beta_5\left\lvert\ln\left(A_i / A_j\right)\right\rvert\right) + \varepsilon$		0.144	-0.864	-0.086	0.565	0.125	-0.140	0.011	80%	92

As shown in Figure 4.10, Model B fits the trend in the data very well. The model effectively explains the actual variation in the cross-correlation from site-pair to site-pair using only distance between the centroids. Residual errors in Figure 4.10 are difficult to interpret. For two sites with a concurrent record length of n, the variance of the estimate of their correlation is $(1-\rho^2)^2/n$. Thus, there are dramatic differences in the standard errors of the estimated cross-correlations depending on the true cross-correlation ρ. Here $70 \leq n < 111$, so that n is relatively constant, whereas ρ may vary widely.

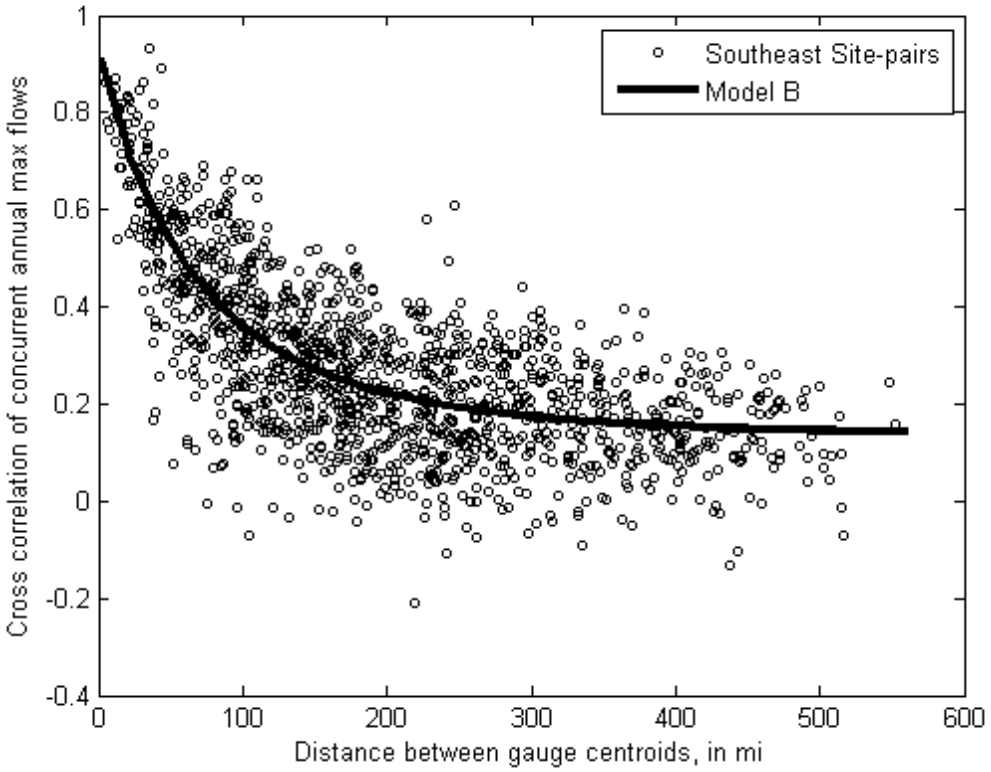

Figure 4.10: Graph of site-to-site cross-correlation ρ_{ij} versus distance between basin centroids, where the Southeastern U.S. site-pairs are those non-redundant pairs with concurrent records greater than 70 years.

A Psuedo ANOVA table was also created to examine the constant model, Model A, as well as Model B more closely. A Pseudo ANOVA table was introduced

in Chapter 2. The table describes the sources of error in the regression and their impact on the total error associated with each model.

Table 4.7: Pseudo ANOVA table for selected models of the cross-correlation regression with n = 1317 site-pairs (redundant sites omitted).

Source	Degrees-of-Freedom			Equations	Sum of squares	
		Model A	Model B		Model A	Model B
Model	k	0	2	$n\left[\sigma_\delta^2(0)-\sigma_\delta^2(k)\right]$	0.0	43
Model Error	n-k-1	1316	1314	$n\sigma_\delta^2(k)$	52	8.9
Sampling Error	n	1317	1317	$\sum_{i=1}^{n} Var(\hat{y}_i)$	18	18
Total	2n-1	2633	2633	$n\sigma_\delta^2(0)+\sum_{i=1}^{n} Var(\hat{y}_i)$	71	71
EVR	$EVR=\frac{\text{SS(sampling error)}}{\text{SS(model error)}}=\frac{\text{tr}[\boldsymbol{\Sigma}(\hat{\mathbf{y}})]}{n\sigma_\delta^2(k)}$				0.35	2.1
$\mathbf{R_\delta^2}$	$R_\delta^2=1-\frac{\sigma_\delta^2(k)}{\sigma_\delta^2(0)}$				0.0%	83%

Table 4.7 compares the errors between the constant model, A, and the selected model of best fit, B employing the distance D between two gauges. This table describes how much of the variation in the observations can be attributed to the regional model, and how much of the residual variation can be attributed to model error and sampling error, respectively. The problem is that one cannot actually resolve what the model errors are because the values of the sampling errors η_i for each *i* are unknown. But, one can describe the total sampling error sum of squares by its mean value, which is $tr[\Sigma(\hat{\mathbf{y}})]$, where $tr[\mathbf{A}]$ is the trace of matrix **A**. And because there are *n* equations, the total variation due to the model error δ for a model with *k* parameters has a mean equal to $n\sigma_\delta^2(k)$. That provides descriptions of two of the three sources of variation.

For a model with no parameters other than the mean, the estimated model error $\sigma_\delta^2(0)$ describes all of the variation in $\hat{y}_i = y_i + \eta_i$ not explained by the sampling errors

η_i. Thus, it should on average equal the actual variation in y due to regression and the variation due to the model errors δ. The TOTAL expected sum of squares variation due to model, model error, and sampling error is described as $n\sigma_\delta^2(0) + tr[\Sigma(\hat{\mathbf{y}})]$. Therefore, an expected sum of squares equal to $n[\sigma_\delta^2(0) - \sigma_\delta^2(k)]$ is attributed to the model. This is called a pseudo ANOVA because the contributions of the three sources of error are estimated or constructed, rather than being determined from the computed residual errors and the observed model predictions, and the impact of correlation among the sampling errors is ignored.

The pseudo ANOVA table also contains the EVR, or error variance ratio, which is a modeling diagnostic used to determine if a simple OLS regression is sufficient or a more sophisticated WLS or GLS analysis is appropriate. EVR is the ratio of the average sampling error variance to the model error variance. An EVR greater than 20%, indicating that the sampling variance is not negligible when compared to the model error variance suggests the need for a WLS or GLS regression analysis.

For Model B, the EVR = 2.1. Because the EVR is much greater than 20%, it suggests the need for WLS or GLS regression to correctly understand the variance structure of the data.

As shown in Table 4.7, for Model B, the model error is a little less than half of the sampling error. Thus, this indicates that Model B is explaining a large portion of the explainable variation in the data. This is illustrated by the Pseudo R_δ^2 statistic which is about 80% for Model B.

Using model B, where Z_{ij} is a function of D_{ij}, the estimator of the cross-correlation ρ_{ij} is

$$\rho_{ij} = \frac{\exp\left[2Z_{ij}\right]-1}{\exp\left[2Z_{ij}\right]+1} \tag{4.20}$$

By construction, $Z_{ij} > 0$, thus as a consequence, $\rho_{ij} > 0$ is always obtained. Furthermore, as D_{ij} increases, Z_{ij} monotonically approaches the intercept $\beta_0 = 0.136$, so that ρ_{ij} will also approach β_0 at large distances. However for $D_{ij} = 0$, $Z_{ij} = 1.57$, so that $\rho_{ij} = 0.935$. Figure 4.11 provides a graph of this ρ_{ij} as a function of D_{ij}.

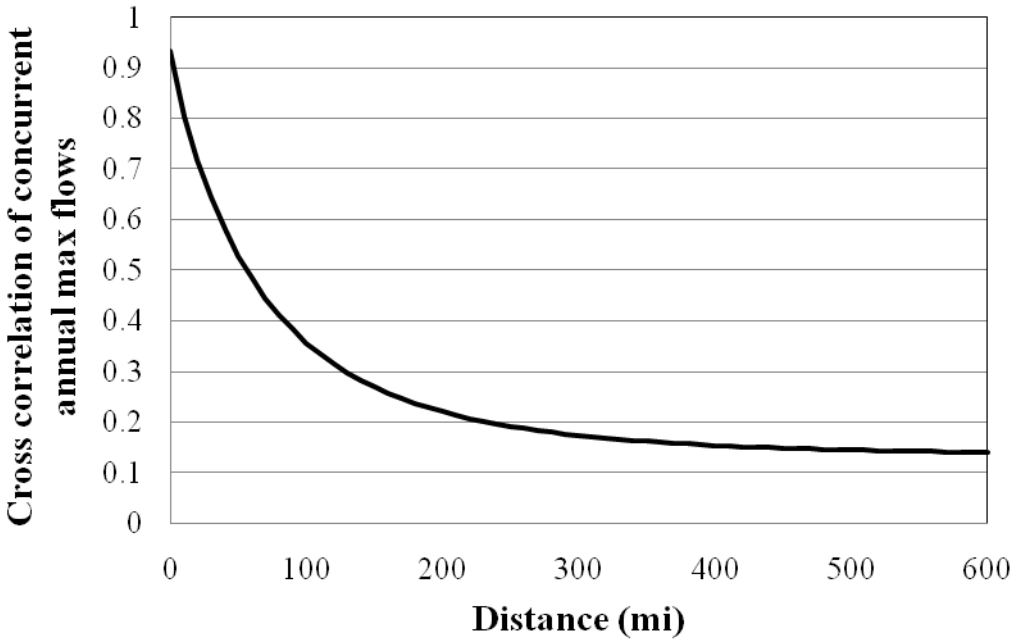

Figure 4.11: Graph of site-to-site cross-correlation versus distance between basin centroids based upon Model B.

In Figure 4.10 there is an outlier with a large correlation. It has the highest cross-correlation of those remaining site-pairs equal to 0.93 and one of the smaller distances between basin centroids equal to 34 mi. It would appear from the size of the residuals and the large cross-correlations that the sites are redundant, however, they were not picked up by the ND-DAR analysis. Thus, this site pair was inspected further. Figure 4.12 depicts the outlines of both basins.

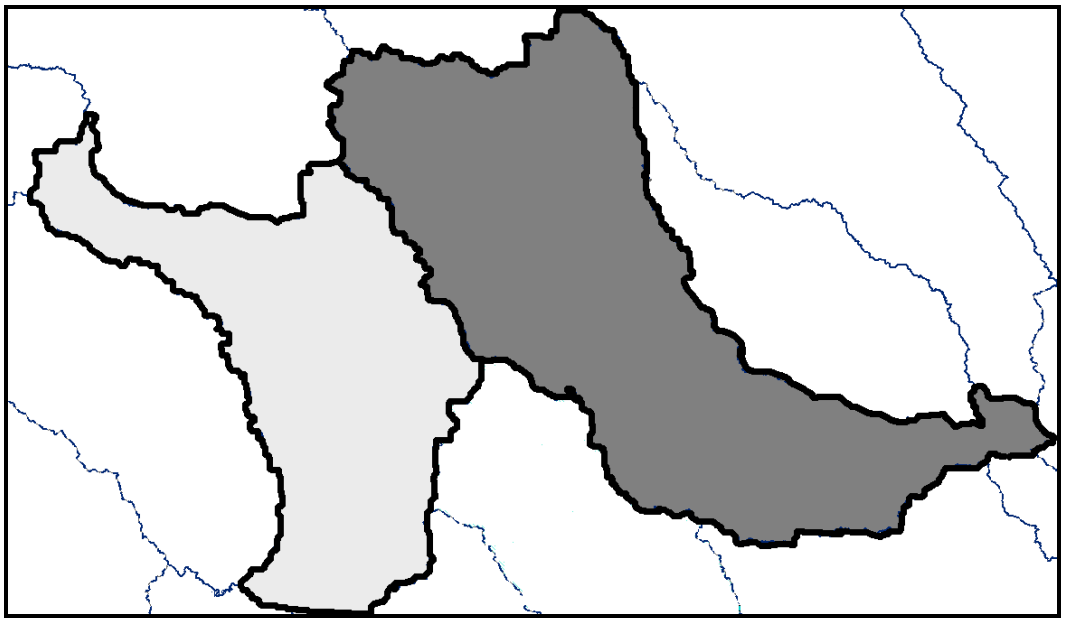

Figure 4.12: Illustration outlining the drainage basins comprising the site-pair with the largest cross-correlation included in the study after screening for redundant sites. The figure was produced using the USGS NHD Geodatabase from http://nhdgeo.usgs.gov/viewer.htm.

The basin on the left, the light gray region, is site number 170, or USGS site 02349605, located on the Flint River in central Georgia. The basin on the right, the dark gray region, is site number 80, or USGS site 02215000, located on the Omculgee River also in central Georgia. Both sites have extremely large drainage areas, 2920 mi^2 and 3760 mi^2, for sites 170 and 80 respectively. Their basin centroids are also located only 34 mi apart, thus the ND for this site pair is equal to 0.59, which is greater than the ND threshold of 0.5. These sites are not a redundant pair as shown by the analysis and also demonstrated in Figure 4.12. This appears to be a case of two gauges located on two larger rivers, located in close proximity. Thus, it is possible that the annual peak flows for both basins are generated from the same large storm events, causing the flows to be highly correlated. These two sites were left in the analysis.

4.5 Bayesian-GLS Regional Skew Regression Results for the Southeastern U.S.

After identifying and screening redundant gage sites, as well as developing a model of the cross-correlations of annual peak flows, Bayesian GLS (B-GLS) regression was employed to identify the best regional skewness model. This section summarizes the B-GLS regression results. Additional details pertaining to the B-GLS methodology can be found in Chapter 2 as well as in Reis *et al*. [2005], Reis [2005], and Gruber *et al*. [2007].

Those sites affected by censored data, in which the flows were less than or greater than the measurement threshold, were not used in the analysis because of corrected skews and lack of available precision. There were 59 sites with censored data in the Southeastern U.S. data set. Also, for those sites which had historic peaks, those historic peaks were not included in the analysis. Thus, in addition to the 92 redundant sites removed from the study, some of which also contained censored data, 55 sites were removed because they contained censored data. Thus, the B-GLS regression was performed on the remaining 342 sites. These 342 sites were not screened for low outliers. The *EMA-Peak* program developed by Tim Cohn that could have given low outliers special treatment was not yet available. Thus, EMA was not used to screen for low outliers before skews were computed for the LP3 distribution. Tables 4.8 and 4.9 below show the breakdown of both the omitted sites, due to redundancy and censored data, as well as those sites remaining in the study by both state and physiographic region.

Table 4.8: Breakdown of sites in Southeastern U.S. study by state after redundant site screening and removal of sites with censored data

	AL	FL	GA	NC	SC	TN	VA	Total
# of Sites Included in Study	18	12	103	107	28	38	36	342
# of Sites Removed from Study due to:								
Redundancy	7	7	27	20	10	12	9	92
Censored Data	0	0	39	0	0	14	2	55
# of Total Sites	25	19	169	127	38	64	47	489

Table 4.9: Breakdown of sites in Southeastern U.S. study by physiographic province after redundant site screening and removal of sites with censored data

	BR	CA	MAC	P	RV	SH	SP	SCP	SA	Total
# of Sites Included in Study	69	0	15	127	36	14	57	15	9	342
# of Sites Removed from Study due to:										
Redundancy	14	1	2	44	8	2	14	5	2	92
Censored Data	3	2	0	12	6	3	22	4	3	55
# of Total Sites	86	3	17	183	50	19	93	24	14	489

Note: Region for each site is chosen according to which region the largest % of basin is contained in.

Table 4.9 shows that there are no sites remaining in the study with the majority of their drainage area located in the Central Appalachian province. However, three sites remain in the study with each between 25% and 40% of their drainage area located in the CA province. However, this does not seem like an adequate amount of data to classify CA as a distinct region. This will be explored deeper after the results of the B-GLS skew regression are presented.

The results from the B-GLS skew regression are shown below. Table 4.10 contains twenty-two single parameter models plus a constant model. A B-GLS regression was performed on all of the available explanatory variables one at a time to evaluate the significance of each parameter by itself.

Table 4.10: Single parameter B-GLS skew regression models for the Southeastern U.S. data set (342 sites). Bayesian standard deviations and plausibility values, as percentages, are presented in parenthesis.

Model	Constant	Physiographic Province/ Basin Parameter	σ_δ^2	Avgerage Sampling Variance	AVP_{new}	R_δ^2
Constant	-0.019 (0.063)	-	0.139 (0.021)	0.0040	0.143	0.0%
Blue Ridge	0.003 (0.063)	0.003 (0.001) (0.6%)	0.134 (0.021)	0.0059	0.140	3.3%
Piedmont	-0.036 (0.063)	-0.003 (0.001) (0.5%)	0.135 (0.021)	0.0056	0.140	2.9%
Sand Hills	-0.021 (0.063)	0.005 (0.002) (0.3%)	0.134 (0.021)	0.0051	0.139	3.6%
Southeastern Plains	-0.025 (0.063)	0.001 (0.00) (21%)	0.139 (0.021)	0.0056	0.144	-0.1%
Middle Atlantic Coastal Plain	-0.020 (0.063)	0.0002 (0.002) (90%)	0.139 (0.021)	0.0052	0.145	-0.6%
Central Appalacians	-0.023 (0.063)	0.008 (0.008) (32%)	0.138 (0.021)	0.0047	0.143	0.1%
Ridge and Valley	-0.018 (0.063)	-0.001 (0.001) (38%)	0.138 (0.021)	0.0054	0.143	0.3%
Southern Coastal Plain	-0.007 (0.063)	-0.006 (0.002) (0.4%)	0.137 (0.021)	0.0058	0.143	0.8%
Southwestern Appaliacians	-0.014 (0.063)	-0.003 (0.002) (15%)	0.139 (0.021)	0.0052	0.144	0.0%
Drainage Area (mi^2)	-0.012 (0.064)	-0.012 (0.019) (51%)	0.139 (0.021)	0.0050	0.144	-0.3%

Table 4.10 (Continued):

Model	Constant	Physiographic Province/ Basin Parameter	σ_δ^2	Avgerage Sampling Variance	AVP_{new}	R_δ^2
Main Channel Slope (ft/mi)	-0.017 (0.063)	0.006 (0.027) (82%)	0.139 (0.021)	0.0053	0.144	-0.4%
Average basin slope (%)	-0.009 (0.063)	0.005 (0.004) (15%)	0.139 (0.021)	0.0063	0.145	0.0%
Main Channel Length (mi)	-0.013 (0.064)	-0.016 (0.031) (60%)	0.139 (0.021)	0.0050	0.144	-0.4%
Basin perimeter length (mi)	-0.014 (0.064)	-0.015 (0.033) (65%)	0.139 (0.021)	0.0050	0.144	-0.4%
Basin shape factor	-0.022 (0.063)	0.006 (0.006) (32%)	0.139 (0.021)	0.0048	0.143	0.0%
Avg basin elev (ft, NAVD88)	-0.007 (0.064)	0.040 (0.035) (26%)	0.138 (0.021)	0.0061	0.144	0.3%
Max basin elev (ft, NAVD88)	-0.011 (0.064)	0.030 (0.034) (37%)	0.139 (0.021)	0.0060	0.145	-0.1%
Avg ann. Precip. in basin (in)	-0.013 (0.063)	0.398 (0.356) (26%)	0.137 (0.021)	0.0060	0.143	1.2%
% basin impervious surfaces	-0.019 (0.063)	-0.001 (0.007) (89%)	0.139 (0.021)	0.0051	0.145	-0.6%
% basin occupied by forests	-0.015 (0.063)	0.001 (0.001) (40%)	0.139 (0.021)	0.0053	0.144	-0.3%
Avg soil drainage index	-0.001 (0.064)	-0.124 (0.058) (3%)	0.136 (0.021)	0.0059	0.142	1.6%
Avg hydrologic soil index	-0.004 (0.063)	-0.186 (0.094) (5%)	0.137 (0.021)	0.0055	0.143	1.1%

As shown in Table 4.10, none of the single parameter models resulted in a major improvement in the model fit as compared to the constant model. In particular, as Figure 4.13 shows, none of the Pseudo R_δ^2 values for the single parameter models are greater than 4%.

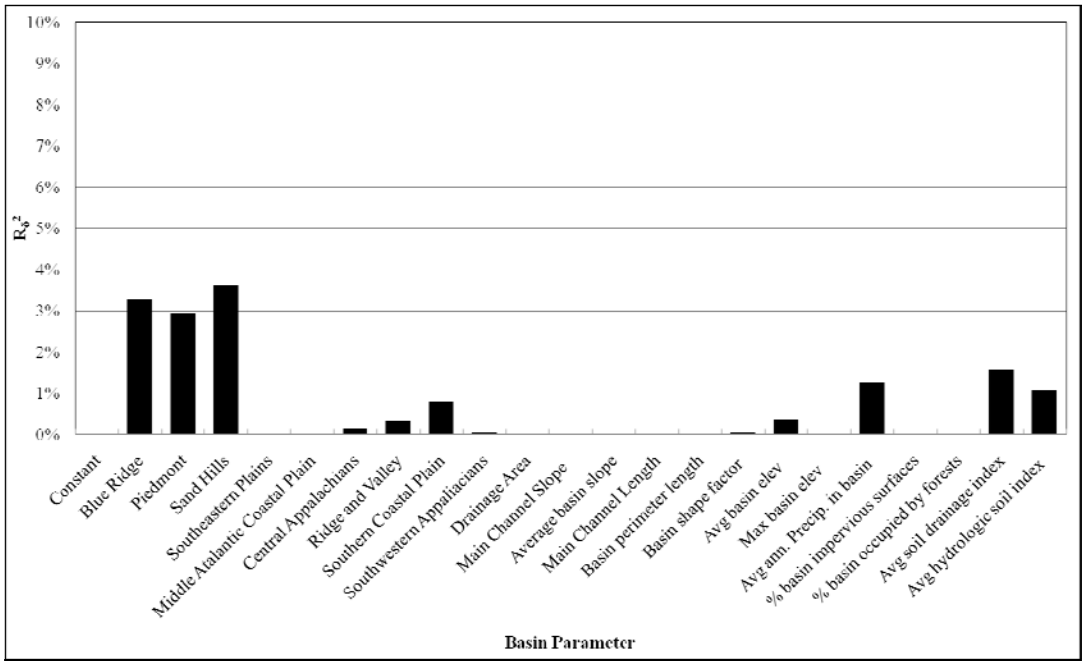

Figure 4.13: Pseudo R_δ^2 values for single parameter Southeastern U.S. B-GLS regional skew regression models.

Based on the results obtained from the single parameter models, all of the physiographic province explanatory variables, as well as average annual precipitation, average basin elevation and average soil drainage index were used to create multi-parameter models. Table 4.11 contains the results.

Table 4.11: Multi-parameter B-GLS skew regression models for the Southeastern U.S. data set (342 sites). Bayesian standard deviations and plausibility values, as percentages, are presented in parenthesis.

Model	Constant	BR	P	SH	SP	MAC	CA	RV	SCP	SA	Avg. Annual Precip.	Avg. soil drainag	Avg. hydrologic soil	σ_δ^2	Average Sampling Variance	AVP_{new}	R_δ^2
Const.	-0.019 (0.063)	-	-	-	-	-	-	-	-	-	-	-	-	0.139 (0.021)	0.0040	0.143	0.0%
A	0.003 (0.063)	0.003 (0.001) (0.6%)	-	-	-	-	-	-	-	-	-	-	-	0.134 (0.021)	0.0059	0.140	3.3%
B	-0.036 (0.063)	-	-0.003 (0.001) (0.5%)	-	-	-	-	-	-	-	-	-	-	0.135 (0.021)	0.0056	0.140	2.9%
C	-0.021 (0.063)	-	-	0.005 (0.002) (0.3%)	-	-	-	-	-	-	-	-	-	0.134 (0.021)	0.0051	0.139	3.6%
D	-0.007 (0.063)	-	-	-	-	-	-	-	-0.006 (0.002) (0.4%)	-	-	-	-	0.137 (0.021)	0.0058	0.143	0.8%
E	-0.013 (0.063)	-	-	-	-	-	-	-	-	-	0.40 (0.36) (26%)	-	-	0.137 (0.021)	0.0060	0.143	1.2%
F	-0.001 (0.064)	-	-	-	-	-	-	-	-	-	-	-0.124 (0.058) (3.1%)	-	0.136 (0.021)	0.0059	0.142	1.6%
G	-0.004 (0.063)	-	-	-	-	-	-	-	-	-	-	-	-0.186 (0.094) (4.7%)	0.137 (0.021)	0.0055	0.143	1.1%
H	0.003 (0.063)	0.003 (0.001) (0.3%)	-	0.005 (0.002) (0.2%)	-	-	-	-	-	-	-	-	-	0.129 (0.034)	0.0070	0.136	6.9%
I	0.003 (0.063)	0.003 (0.001) (1.3%)	-	0.005 (0.002) (0.2%)	-	-	-	-	-	-	0.001 (0.007) (92%)	-	-	0.129 (0.020)	0.0087	0.138	6.6%
J	-0.010 (0.064)	0.003 (0.001) (2.3%)	-0.001 (0.001) (21%)	0.005 (0.002) (0.7%)	-	-	-	-	-	-	-	-	-	0.129 (0.020)	0.0088	0.138	6.9%
K	-0.003 (0.064)	0.002 (0.001) (5.3%)	-0.002 (0.001) (9.9%)	0.004 (0.002) (1.4%)	-	-	-	-	-0.006 (0.002) (0.6%)	-	-	-	-	0.128 (0.020)	0.0105	0.139	7.4%
L	0.001 (0.064)	-	-	-	-	-	-	-	-	-	-	-0.087 (0.082) (29%)	-0.086 (0.133) (52%)	0.137 (0.021)	0.0070	0.144	1.3%
M	-0.008 (0.064)	-0.055 (0.067) (41%)	-0.059 (0.067) (38%)	-0.053 (0.067) (43%)	-0.057 (0.067) (40%)	-0.057 (0.067) (39%)	-0.046 (0.067) (49%)	-0.059 (0.067) (38%)	-0.063 (0.067) (34%)	-0.062 (0.068) (36%)	-	-	-	0.128 (0.020)	0.0159	0.143	50%
N	-0.007 (0.064)	-0.056 (0.068) (41%)	-0.060 (0.068) (37%)	-0.054 (0.068) (42%)	-0.057 (0.068) (40%)	-0.057 (0.068) (40%)	-0.047 (0.068) (49%)	-0.060 (0.068) (37%)	-0.063 (0.068) (35%)	-0.063 (0.068) (36%)	-	-0.044 (0.102) (66%)	-0.003 (0.145) (98%)	0.129 (0.020)	0.0182	0.147	50%

Table 4.11 first presents the constant model and then subsequently presents the best single parameter models and then the best multi-parameter models. However, most of the regression coefficients in the multi-parameter models have Bayesian Plausibility Values greater than 10%, indicating that zero is a plausible value. This

implies that these regression coefficients are not statistically different from zero. In comparing multi-parameter models with statistically significant coefficients to the single parameter models, it is evident that the multi-parameter models do have larger Pseudo R_{δ}^2 values. However, the Pseudo R_{δ}^2 values are still under 10% indicating that the use of several explanatory variables does not result in a major improvement in the fit as compared to the constant model. Alternatively, the addition of explanatory variables to the model does increase the complexity of the model. Thus, the regional skew model chosen for use is the constant model.

The constant model has a regional skew, $\hat{\gamma} = -0.019$ with $\text{AVP}_{new} = 0.14$. Setting this AVP_{new} equal to the mean square error of a biased sample skewness estimator yields an effective record length (ERL) of about 39 years. Because the regional skew regression model is a constant, the variance of prediction at a new site is also constant.

Table 4.12 presents a Pseudo ANOVA table for the B-GLS skew regressions. The Pseudo ANOVA table describes how much of the variation in the sample can be attributed to the model, and how much to model error and sampling error, respectively. This table compares the constant model with Model H from Table 4.11. Model H was chosen for comparison because all of its explanatory variables are significant at the 5% level and it has the smallest model error variance and largest Pseudo R_{δ}^2.

As shown in Table 4.12, both the constant model and model H have sampling error variances larger than their model error variances. Also, it is evident that the addition of the two explanatory variables in Model H, Blue Ridge and Middle Atlantic Coastal Plain, do not result in a major decrease in the model error. This indicates that these explanatory variables are not helping to improve the model fit. The EVR is greater than 1, clearly indicating that either a WLS or GLS analysis should be

employed as opposed to an OLS analysis to correctly understand the error structure of the data. Moreover, because the MBV exceeds 5, a GLS analysis is clearly appropriate; a WLS analysis would overestimate the precision of b_0, the single parameter of the constant model. Thus, to neglect the cross-correlation of the skewness estimators would have resulted in a major distortion in the estimated precision of the overall average skew and possibly the differences among physiographic provinces. Similarly, provincial variables are very much like indicator variables that have a constant value in each province, so their precision is likely to be misrepresented as well.

Table 4.12: Pseudo ANOVA table for B-GLS Southeastern U.S. regional skew regression

Source	**Degrees-of-Freedom**			**Equations**	**Sum of squares**	
		Constant	H		Constant	H
Model	k	0	2	$n\left[\sigma_\delta^2(0)-\sigma_\delta^2(k)\right]$	0.0	3.3
Model Error	n-k-1	341	339	$n\sigma_\delta^2(k)$	47	44
Sampling Error	n	342	342	$\sum_{i=1}^{n} Var(\hat{y}_i)$	59	59
Total	2n-1	683	683	$n\sigma_\delta^2(0)+\sum_{i=1}^{n} Var(\hat{y}_i)$	107	107
EVR	$EVR=\frac{\text{SS(sampling error)}}{\text{SS(model error)}}=\frac{\text{tr}[\boldsymbol{\Sigma}(\hat{\mathbf{y}})]}{n\sigma_\delta^2(k)}$				1.3	1.3
MBV	$MBV=\frac{Var[b_0^{WLS}\mid GLS\,analysis]}{Var[b_0^{WLS}\mid WLS\,analysis]}=\frac{w^T\Lambda w}{N}$ where $w_i=\frac{1}{\sqrt{\Lambda_{ii}}}$				5.4	5.5
$\mathbf{R_\delta^2}$	$R_\delta^2=1-\frac{\sigma_\delta^2(k)}{\sigma_\delta^2(0)}$				0.0%	6.9%

As shown in Table 4.11, the posterior mean of the model error variance σ_δ^2 of the constant model is 0.14 with a standard error of 0.021. The standard error of the model error variance is much smaller than those obtained in previous B-GLS studies

due to the large sample size employed here. The constant model has an average variance of prediction (AVP) at a new site equal to 0.14, which is the same as the model error variance. This is a significant improvement over the Bulletin 17B skew map value with a mean square error of 0.30 [IACWD 1982] with an effective record length of 17 years. Use of province and precipitation made very modest improvements in precision, and do not seem worthwhile given the additional complexity that they would introduce.

When employing the constant model, when a site is included in a regression analysis (old site), there is a small reduction in the variance of estimation at that site over the variance of prediction at a new site. Because of the large sample size employed here, the difference is very small and frequency analyses can just use a MSE of 0.14 in all cases.

4.6 Sensitivity Analysis for B-GLS Southeastern U.S. Skew Models

This section offers a diagnostic analysis of the best B-GLS regional skew models for the Southeastern U.S. study. Figure 4.14 displays the leverage and influence values for the B-GLS constant model. The 21 sites included in the figure have high leverage, high statistical leverage, high influence, and/or high σ-influence. The sites are ordered, starting from the left, by decreasing influence, as it identifies those sites that had a large impact on the analysis. Differences in leverage values reflect the variation in record lengths among sites.

Site 26 has the highest σ-influence value, implying it has a larger impact on the model error variance, while also having high influence. However, Site 26 does not have high leverage values, and thus, is not likely to have a large impact on the estimated regression constant in this model. Site 26 has a log-space at-site skewness value of -2.04, a record length of 38 years, a drainage area of 126 mi^2, and the second

largest residual, -2.3. If Site 26 is removed from the Case 1 [Constant (all sites)] model as a test, its large impact on the model error variance is apparent. As expected, σ_δ^2 of Case 2 [Constant (w/o site 26)] model decreased from 0.14 to 0.13, as shown in Table 4.13.

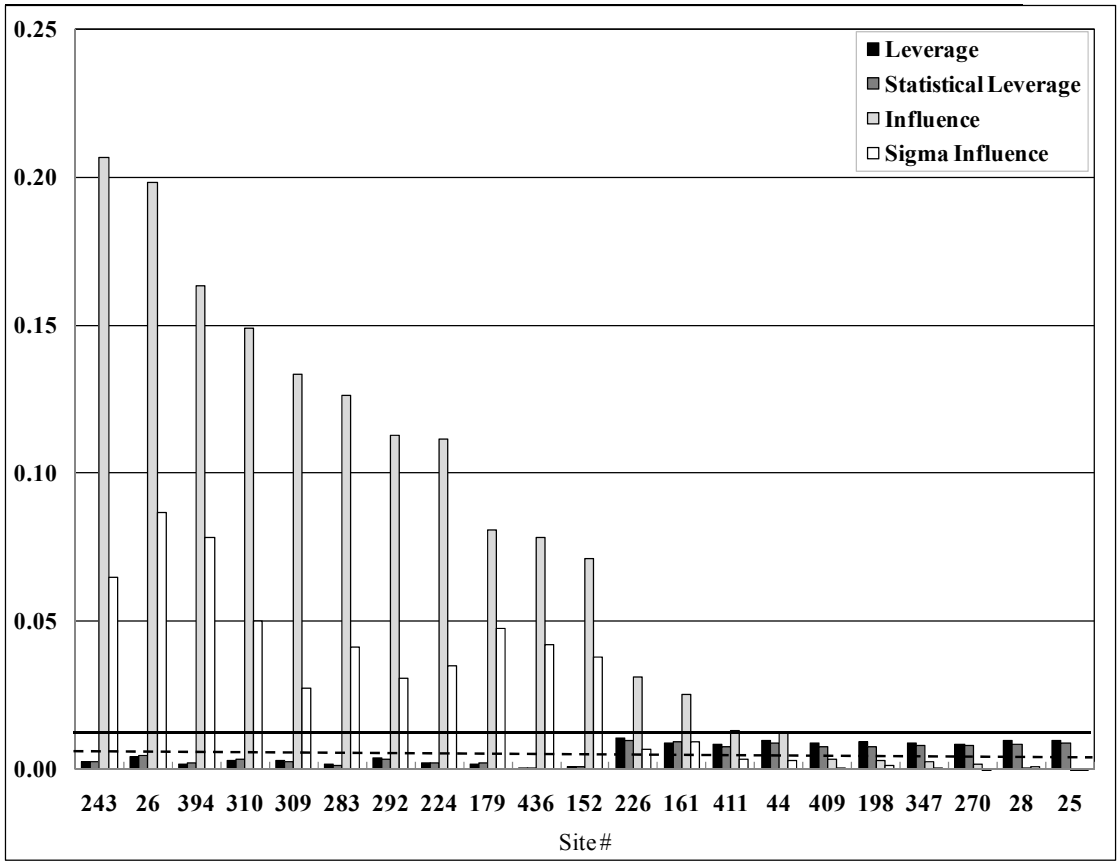

Figure 4.14: Regression Diagnostics: Leverage and influence for the Southeastern U.S. B-GLS Constant Model. The solid line represents the threshold for high influence and σ-influence, while the dotted line represents the threshold for high leverage and statistical leverage.

Table 4.13: Sensitivity Analysis for the Constant Skew Regression Model using the Southeastern U.S. data set (342 sites); standard errors are reported in parenthesis.

Model	Constant	σ_δ^2	Avgerage Sampling Variance	AVP_{new}	R_δ^2
Case 1: Constant (all sites)	-0.019 (0.063)	0.14 (0.021)	0.0040	0.14	0.0%
Case 2: Constant (w/o site 26)	-0.011 (0.063)	0.13 (0.020)	0.0039	0.13	0.0%
Case 3: Constant (w/o site 226)	-0.024 (0.063)	0.14 (0.021)	0.0040	0.14	0.0%

As an experiment, Site 226 was also removed from the Case 1 [Constant (all sites)] model as a test. Site 226 has a log-space at-site skewness value of 0.58, a record length of 70 years, a drainage area of 2813 mi^2, and a residual of 0.65. Like Site 26, Site 226 has an influence value larger than the threshold, however that is the only similarity. Site 226 also has high leverage and high statistical leverage, with a σ-influence that falls below the threshold. As shown in Table 4.13, when Site 226 is removed, neither the σ_δ^2 or the AVP_{new} change from those generated for the Case 1 [Constant (all sites)] model. This is due to the fact that Site 226 does not have high σ-influence. Site 226 does have an impact on the regression parameters due to the fact it has 70 years of record, which is a relatively long record in the Southeastern U.S. data set.

Table 4.14 contains the pseudo ANOVA results for the Constant model as well as the two sensitivity analysis. The pseudo ANOVA table clearly demonstrates that for all cases in the Southeastern U.S., the sampling error is larger than the model error. As shown in Table 4.14, the model error for the Case 2 [Constant (w/o site 26)] model is slightly less than those results obtained by both the Case 1 and Case 3 models.

Table 4.14: Pseudo ANOVA table for the Southeastern U.S. Constant Skew Regression Models presented in Table 4.13.

Source	Degrees-of-Freedom			Sum of squares		
		Case 1	Cases 2,3	Case 1 (all sites)	Case 2 (w/o site 26)	Case 3 (w/o site 226)
Model	k	0	0	0.0	0.0	0.0
Model Error	n-k-1	341	340	47	44	47
Sampling Error	n	342	341	59	59	59
Total	2n-1	683	681	107	103	107
EVR				1.3	1.4	1.3
MBV				5.4	5.5	5.4
$\mathbf{R_\delta^2}$				0.0%	0.0%	0.0%

To better understand the behavior of the leverage and influence diagnostics, the results from the *Constant_Model* were compared to the results from the *Central_Appalachian_Model* in Table 4.10. The *Central_Appalachian_Model* for regional skew is

$$\hat{\gamma} = -0.023 + 0.008\mathbf{X}_{CA} \tag{4.21}$$

where $\mathbf{X}_{CA}$ is the centered variable representing the percent of each basin's drainage area that lies in the CA province. The regression parameter in front of $\mathbf{X}_{CA}$ has a Bayesian Plausibility value of 32%, which is not statistically different from zero. Due to the redundant site screening, as well as, the removal of those sites with censored data, there are only three sites remaining in the study with drainage area in the CA province. Introducing a province with so little supporting data seems unwise. The leverage and perhaps influence metrics should support such a recommendation.

Table 4.15: Leverage values for the three sites in the study representing Central Appalachian province for *Constant_Model* (Constant), as well as the C*entral_Appalachian_Model* (CA).*

			% of Basin in Province			Leverage		Statistical Leverage		Residual	
Site #	**USGS #**	**POR**	CA	RV	SA	Constant	CA	Constant	CA	Constant	CA
409	03528000	86	28.8	71.2	0	*0.009*	*0.253*	*0.008*	*0.118*	0.200	-0.029
411	03532000	73	35.4	64.8	0	*0.008*	*0.359*	*0.008*	*0.173*	0.416	0.134
420	03538500	48	39.2	0	60.8	0.005	*0.388*	0.005	*0.208*	0.236	-0.078

* The italicized diagnostic values are values which are greater than their respective threshold: *Constant_Model* leverage thresholds = 0.006, *CA_Model* leverage thresholds = 0.012.

Table 4.16: Influence values for the three sites in the study representing Central Appalachian province for *Constant_Model* (Constant), as well as the C*entral_Appalachian_Model* (CA).*

			% of Basin in Province			Influence		σ-Influence		Residual	
Site #	**USGS #**	**POR**	CA	RV	SA	Constant	CA	Constant	CA	Constant	CA
409	03528000	86	28.8	71.2	0	0.004	0.001	0.000	0.000	0.200	-0.029
411	03532000	73	35.4	64.8	0	*0.013*	*0.036*	0.004	0.000	0.416	0.134
420	03538500	48	39.2	0	60.8	0.003	0.010	0.001	0.000	0.236	-0.078

* The italicized diagnostic values are values which are greater than their respective threshold:*Constant_Model* influence thresholds = 0.012, *CA_Model* influence thresholds = 0.012.

Tables 4.15 and 4.16 provide the leverage and influence values for both the *Constant_Model* and the *CA_Model* for the three sites with drainage area in the CA province. The three sites in are all located in Tennessee and all contain between 25% and 40% of their basin drainage area in the Central Appalachian province. There are no other sites that have any of their area in the Central Appalachian province. All three sites have long record lengths, especially sites 409 and 411. Across all sites, leverage values for the Constant Model sum to one, while the leverage values for the CA Model sum to two; this is due to the fact that each variable in the regression adds a unit of leverage to the model [Tasker and Stedinger, 1989, eqn 23]. In the Constant Model, sites 409 and 411 have leverage values just greater than the threshold. However, in

the CA Model all three sites have leverage values 200-300 times greater than the threshold, and in fact their values appears to sum to 1. So the leverage values for these sites are in fact stupendous. The three sites in Table 4.15 are the only three sites with leverage values greater than the threshold in the CA Model. Thus, they are capturing almost all of the leverage for the second variable, $\mathbf{X}_{CA}$. Site 420 has the largest leverage value as it has the largest percentage of its basin in the CA province and the smallest sample size. Leverage depends only the weighting matrix Λ and the values of the $\mathbf{X}$ variances, and does not depend on the value of the regression coefficients β, so the fact that the coefficient of $\mathbf{X}_{CA}$ in eqn, 4.21 is small is not particularly important.

The statistical leverage value considers the effect on the regression due to the statistical variation in each residual. For the *CA_Model*, all three sites have statistical leverage values great than the threshold. Site 420 has the smallest record length and thus the largest statistical leverage. This is due to the fact that the shorter record length allows for more statistical fluctuations or errors.

The influence metric is helpful in highlighting those sites which have an unusual impact on the regression analysis, which is a combination of an unusually large residual and large leverage. For both the Constant Model and the CA Model, site 411 is the only site with influence values above the thresholds, as shown in Table 4.16. The residuals of site 411 are the largest of the three sites presented in the Table 4.16. This is due to the fact that the at-site skew values (site 409 = 0.17, site 411 = 0.37, site 420 = 0.19) of the three sites are small and relatively similar, but is largest for site 411. However, if the at-site skew values had been larger and different from each other, their residuals would have been larger, and thus their influence values would be much larger.

As shown in Table 4.16, the σ-Influence values for both models are all about zero as both models are fitting the three sites fairly well. This is due to the fact that all

3 residuals are relatively small, and thus none of the sites are having an unusual impact on the estimated model error variance. The residuals these three sites are smaller with the CA Model because the CA Model has an extra parameter to specifically fit those three sites.

Thus, it seems clear that the incredibly large leverage values for the CA Model signal that there is not sufficient data to fit this model. This is consistent with the large Bayesian Plausibility value on the $\mathbf{X}_{CA}$ regression parameter as presented in Table 4.10 indicating the regression coefficient was not significant.

4.7 Comparison of Methods: B-GLS Regional Skew versus Bulletin 17B Regional Skew Map

As discussed in the Introduction, the skew map is the current technique for estimating regional skew as developed in *Bulletin 17B*. The new methods proposed and carried out above aim to develop a more accurate and reliable estimate of the regional skew. In this section, the two methods, referred to as *Bulletin 17B* Skew Map and B-GLS Regional Skew Regression, respectively, are compared in terms of their effects on flood frequency.

In order to compare the methods, the methods used in *Bulletin 17B* are first reviewed. As stated previously, *Bulletin 17B* recommends the use of the Log-Pearson Type III distribution to fit annual maximum flood peaks. Equation 1 in *Bulletin 17B*, reproduced below as Equation 4.21, is used to estimate the peak annual max flow for a specified exceedence probability

$$Log_{10}Q = \bar{X} + KS \tag{4.21}$$

in which, $Log_{10}Q$ is the log of the flow corresponding to a specified exceedence probability, $\bar{X}$ is the mean of the logarithms of the annual maximum flows, S is the standard deviation of the annual max flows, and K is a function of the skew coefficient

and the exceedence probability of interest which can be found in tables provided in Appendix 3 of *Bulletin 17B*.

As stated, *K* is a function of not only the exceedence probability of interest, but also the skew coefficient. *Bulletin 17B* recommends using a weighted skew coefficient, denoted G_w, which combines the at-site station skew with the regional skew. G_w is calculated as

$$G_W = \frac{MSE_{\bar{G}} * G + MSE_G * \bar{G}}{MSE_{\bar{G}} + MSE_G} \tag{4.22}$$

where G is the at-site station skew, $\bar{G}$ is the regional skew, MSE_G is the mean square error of the at-site station skew, and $MSE_{\bar{G}}$ is the mean square error of the regional skew. *Bulletin 17B* provides the following set of equations to calculate MSE_G.

$$MSE_G = 10^{\left[A - B\left(Log_{10}(N/10)\right)\right]} \tag{4.23a}$$

$$\text{where } A = \begin{cases} -0.33 + 0.08|G| & \text{if } |G| \le 0.90 \\ -0.52 + 0.30|G| & \text{if } |G| > 0.90 \end{cases} \tag{4.23b}$$

$$B = \begin{cases} 0.94 - 0.26|G| & \text{if } |G| \le 1.50 \\ 0.55 & \text{if } |G| > 1.50 \end{cases} \tag{4.23c}$$

where N is the number of years of peak flows in the at-site record. The $MSE_{\bar{G}}$ is constant for the entire skew map and equal to 0.302. Griffis *et al*. [2004] derived a more accurate and smooth approximation of MSE_G, however for the purposes of the following comparison the formulation stated in *Bulletin 17B*, presented above, was applied.

Thus, it is easy to see from Equation 4.22, that not only are the estimates of both the at-site and the regional skew important, but mean square error of both terms is also important. The estimated skew with the smaller error will be weighted more heavily when calculating the weighted skew.

Now that the framework has been established for calculating flood frequencies, the two methods, *Bulletin 17B* Skew Map and B-GLS Regional Skew Regression, can be compared. It is important to note that these two methods only differ in their estimation procedure of the regional skew coefficient. In order to compare the two methods, three sites have been chosen from the Southeastern U.S. data set.

The first site is USGS site # 02049700, or site index # 451, located in Virginia. It has a drainage are of 8.4 mi^2, 27 years of record, and an at-site skew of 0.331 with a MSE_G equal to 0.213. According the regional skew map in *Bulletin 17B*, this site is located in an area with a regional skew of 0.7. This is the largest regional skew the Bulletin 17B skew map delineates for the Southeastern U.S. region. This is compared the regional skew of -0.019 calculated by the B-GLS Regional Skew Regression method. Table 4.17 compares the flood frequency estimates for the 2-year, 50-year and 100-year floods at this site using both methods.

Table 4.17: Regional and Weighted Skew Estimates and Flood Frequency Estimates for USGS Site # 02049700 with at-site $G = 0.331$, $MSE_G = 0.213$ and $N = 27$ years. The column in grey represents the true regional skew $\overline{G} = 0.7$, as denoted by the Bulletin 17B skew map.

	Case 0.00	**Case 0.35**	**Case 0.70**	**Case B-GLS**
	Regional and Weighted Log Skew Estimates			
$\overline{G}$	0.000	0.350	0.700	-0.019
$MSE_{\overline{G}}$	0.302	0.302	0.302	0.140
$\mathbf{G_W}$	0.194	0.339	0.484	0.120
	Flood Frequency Estimates (in cfs)			
$\mathbf{Q_2}$	129	128	126	130
$\mathbf{Q_{50}}$	365	378	392	358
$\mathbf{Q_{100}}$	423	444	466	412

As shown in Table 4.17, the regional skew for USGS Site #02049700 according to *Bulletin 17B* is 0.7 with a $MSE_{\overline{G}} = 0.302$, while the B-GLS estimation

technique predicts the regional skew to be -0.019 with a $MSE_{\bar{G}} = 0.140$. This is a 103% decrease in regional skew and a 54% decrease in $MSE_{\bar{G}}$. This in turn, results in an over prediction of 54 cfs or 13% for the 100-year flood as estimated by *Bulletin 17B*. It is important to note in this example the difference in effects by the two methods. When using the *Bulletin 17B* methods to estimate flood quantiles, the larger weight is placed on the at-site skew when calculating the weighted skew, as it has the smaller MSE (at-site $MSE_G = 0.213$ compared to regional $MSE_{\bar{G}} = 0.302$). It is important to note that only a short record, 27 years, is available at the site. However, when using the B-GLS estimation technique the larger weight is placed on the regional skew when calculating the weighted skew, as it has the smaller MSE (regional skew $MSE_{\bar{G}} = 0.140$ compared to at-site skew $MSE_G = 0.213$). This is the cause for the large difference in the weighted skews calculated by the two methods.

Table 4.17 also contains two other columns under the Bulletin 17B estimation technique heading. These columns are hypothetical. They are there to allow for comparisons supposing that this exact site, with the identical basin characteristics and at-site skew value, was located in a different area in the Southeastern U.S. which had a different regional skew according the *Bulletin 17B* skew map. The same calculations are performed to determine what the weighted skew would be as well as they predicted 2-year, 50-year and 100-year flood events. This allows for a more thorough comparison with the results from the B-GLS estimation technique. As expected, the closer the regional skew value estimated from the *Bulletin 17B* map is to the B-GLS regional skew value the closer the predictions of the weighted skew and flood quantiles. Table 4.18 and 4.19 provide two more examples of the effect of the regional skew prediction on flood quantile estimation using two other sites from the Southeastern U.S. skew study with different at-site skews.

Table 4.18: Regional and Weighted Skew Estimates and Flood Frequency Estimates for USGS Site # 02341900 with at-site $G = 0.869$, $MSE_G = 0.263$ and $N = 28$ years. The column in grey represents the true regional skew $\overline{G} = 0$, as denoted by the Bulletin 17B skew map.

	Case 0.00	**Case 0.35**	**Case 0.70**	**Case B-GLS**
	Regional and Weighted Log Skew Estimates			
$\overline{G}$	0.000	0.350	0.700	-0.019
$MSE_{\overline{G}}$	0.302	0.302	0.302	0.140
$\mathbf{G_W}$	0.464	0.627	0.790	0.289
	Flood Frequency Estimates (in cfs)			
$\mathbf{Q_2}$	867	849	408	888
$\mathbf{Q_{50}}$	5654	5995	6367	5255
$\mathbf{Q_{100}}$	7527	8231	8982	6825

Table 4.19: Regional and Weighted Skew Estimates and Flood Frequency Estimates for USGS Site # 02318700 with at-site $G = 0.052$, $MSE_G = 0.188$ and $N = 27$ years. The column in grey represents the true regional skew $\overline{G} = 0$, as denoted by the Bulletin 17B skew map.

	Case 0.00	**Case 0.35**	**Case 0.70**	**Case B-GLS**
	Regional and Weighted Log Skew Estimates			
$\overline{G}$	0.000	0.350	0.700	-0.019
$MSE_{\overline{G}}$	0.302	0.302	0.302	0.140
$\mathbf{G_W}$	-0.032	0.102	0.237	-0.033
	Flood Frequency Estimates (in cfs)			
$\mathbf{Q_2}$	3695	3627	3561	3696
$\mathbf{Q_{50}}$	20008	21239	22512	19999
$\mathbf{Q_{100}}$	24967	27101	29386	24952

When comparing Tables 4.17, 4.18, and 4.19 the major differences are the at-site skews of the three sites. The site in Table 4.17 has an at-site skew of 0.331, while the site in Table 4.18 has an at-site skew of 0.869 and the site in Table 4.19 has an at-site skew of 0.052. The at-site mean squared errors of three sites are smaller than the $MSE_{\overline{G}}$ of the regional *Bulletin 17B* skew while at the same time greater than the

$MSE_{\bar{G}}$ of the regional B-GLS skew. Thus for these sites, the difference in regional skew and its MSE will have a larger impact on the weighted skew. It is for these sites that the B-GLS regional skew will be weighted more heavily than the at-site skew. However, when the *Bulletin 17B* regional skew is used, it will be weighted less than the at-site skew. Thus, for these sites with short record lengths, that the B-GLS estimation technique for regional skew will allow for the regional skew to be weighted more heavily, and thus provide a more accurate flood estimate.

4.8 Conclusions

The regional regression framework developed by Reis [2005] and Reis *et al*. [2005], with the regression diagnostic statistics discussed in Chapter 2 were used to develop a regional skewness estimator for the Southeastern United States, nominally Georgia, North Carolina, and South Carolina. Based upon a Bayesian Generalized Least Squares analysis of the selected 342 stations, a constant generalized regional skew model was selected for the Southeastern U.S. Region described by the equation:

$$\hat{\gamma} = -0.019$$

with MSE = 0.14. More complicated models were evaluated, but resulted in very modest improvements in accuracy. Thus, they did not seem justified in view of the increased complexity. The constant model with a MSE of 0.14 is a definite improvement over the *Bulletin 17B* skew map which reported a MSE of 0.302. Much of the difference occurs because the GLS analysis correctly reflects both the difference between the sampling error in at-site skew coefficient estimators and the precision of the regional model.

APPENDIX A

SOUTHEASTERN U.S. STREAM FLOW GAUGE SITES

This appendix contains the 489 sites used in the Southeastern U.S. regional flood skewness estimate. Table A contains the USGS site number, site index, years of record, sample at-site log skew, drainage basin area, centroid location, and physiographic province for each of the 489 gauge stations in the study.

Table A: The 489 peak stream flow gauge sites and their basin characteristics used in the Southeastern U.S. regional skew study.

USGS Hydrologic Unit Code	Site Index #	State	Years of Record	Sample at-site log skew	Drainage Area (mi^2)	Basin Centroid Location		Physigraphic Provience (underline indicates <70% of basin in region)
						Lat	Lon	
02342500	1	AL	58	-0.102	321	-85.2	32.5	Southeastern Plains
02342933	2	AL	41	0.451	112	-85.4	32.0	Southeastern Plains
02343275	3	AL	32	0.487	48.2	-85.2	31.6	Southeastern Plains
02343300	4	AL	37	0.893	146	-85.2	31.5	Southeastern Plains
02360000	5	AL	44	0.640	86.6	-85.5	31.8	Southeastern Plains
02360275	6	AL	29	0.422	102	-85.6	31.6	Southeastern Plains
02361000	7	AL	76	0.599	687	-85.5	31.6	Southeastern Plains
02363000	8	AL	75	0.068	499	-85.7	31.9	Southeastern Plains
02398300[r]	9	AL	77	0.270	367	-85.4	34.5	Ridge and Valley
02399000	10	AL	38	-0.754	126	-85.5	34.5	Southwestern Appalachians
02399200[r]	11	AL	49	0.041	200	-85.6	34.5	Southwestern Appalachians
02400100	12	AL	41	-0.864	254	-85.5	33.9	Ridge and Valley
02401000	13	AL	68	-0.276	182	-85.8	34.4	Ridge and Valley
02401500	14	AL	37	0.565	252	-86.3	33.9	Ridge and Valley
02404000	15	AL	55	-0.489	278	-85.7	33.7	Ridge and Valley
02404400[r]	16	AL	42	-0.189	481	-85.8	33.6	Ridge and Valley
02404500[r]	17	AL	35	-0.313	499	-85.8	33.6	Ridge and Valley
02412000	18	AL	53	0.430	448	-85.2	33.8	Piedmont
02412500[r]	19	AL	51	-0.527	793	-85.4	33.7	Piedmont
02413300[r]	20	AL	30	-0.469	406	-85.2	33.6	Piedmont
02413475	21	AL	25	-1.28	47.1	-85.4	33.3	Piedmont
02413500[r]	22	AL	53	-0.788	591	-85.3	33.5	Piedmont
02415000	23	AL	41	-0.447	189	-85.9	33.2	Piedmont
02419000	24	AL	75	-0.390	336	-85.5	32.5	Southeastern Plains
03574500	25	AL	69	0.012	321	-86.2	34.9	Southwestern Appalachians
02229000	26	FL	38	-2.04	126	-82.4	30.3	Southern Coastal Plain
02230000	27	FL	27	-0.411	20.8	-82.1	30.2	Southern Coastal Plain
02231000	28	FL	78	0.065	675	-82.3	30.4	Southern Coastal Plain
02231250	29	FL	29	0.423	20.0	-81.9	30.7	Southern Coastal Plain
02231280	30	FL	40	0.219	29.5	-81.9	30.5	Southern Coastal Plain
02246300	31	FL	39	0.081	31.5	-81.8	30.3	Southern Coastal Plain
02246828	32	FL	28	-0.364	27.5	-81.5	30.3	Southern Coastal Plain
02315000[r]	33	FL	27	-1.34	2097	-82.6	30.9	Southern Coastal Plain
02315500[r]	34	FL	81	-0.634	2427	-82.6	30.8	Southern Coastal Plain
02315550[r]	35	FL	35	-0.650	2647	-82.6	30.8	Southern Coastal Plain
02319000[r]	36	FL	75	-0.043	2147	-83.5	31.1	Southeastern Plains
02319500[r]	37	FL	79	-0.068	6985	-83.1	31.0	Southeastern Plains
02326500	38	FL	40	-1.36	730	-83.7	30.6	Southeastern Plains
02326900	39	FL	46	-0.280	531	-84.0	30.5	Southeastern Plains

Table A (Continued)

USGS Hydrologic Unit Code	Site Index #	State	Years of Record	Sample at-site log skew	Drainage Area (mi^2)	Basin Centroid Location		Physigraphic Provience (underline indicates <70% of basin in region)
						Lat	Lon	
02329000[r]	40	FL	81	-0.180	1146	-84.1	31.0	Southeastern Plains
02329500	41	FL	40	0.292	234	-84.5	30.7	Southeastern Plains
02329600[cr]	42	FL	41	0.349	302	-84.5	30.7	Southeastern Plains
02330100	43	FL	53	0.139	127	-84.8	30.5	Southeastern Plains
02359000	44	FL	72	0.351	839	-85.3	30.9	Southeastern Plains
02177000	45	GA	81	0.529	203	-83.2	35.0	Blue Ridge
02178400	46	GA	42	0.011	58.3	-83.5	35.0	Blue Ridge
02182000	47	GA	51	-0.070	32.5	-83.4	34.7	Piedmont
02188500	48	GA	35	-0.492	38.5	-83.1	34.3	Piedmont
02191200	49	GA	29	-0.083	60.7	-83.5	34.4	Piedmont
02191300	50	GA	108	-0.104	756	-83.3	34.3	Piedmont
02191930[c]	51	GA	43	-0.822	5.24	-83.1	33.8	Piedmont
02191970	52	GA	27	-0.145	1.89	-83.0	33.9	Piedmont
02192000[r]	53	GA	75	-0.368	1418	-83.2	34.2	Piedmont
02192300	54	GA	29	-0.623	0.96	-82.8	33.8	Piedmont
02193500	55	GA	39	-0.546	292	-82.9	33.7	Piedmont
02197520	56	GA	25	-1.01	56.1	-82.6	33.4	Piedmont
02197830[r]	57	GA	27	0.689	473	-82.3	33.3	Southeastern Plains
02198000	58	GA	69	0.137	646	-82.2	33.2	Southeastern Plains
02200100	59	GA	25	0.071	54.0	-82.9	33.3	Sand Hills
02200400	60	GA	27	-0.212	191	-82.6	33.3	Sand Hills
02200500[r]	61	GA	31	-0.678	806	-82.7	33.3	Piedmont
02200900	62	GA	26	-0.019	95.5	-82.4	33.1	Southeastern Plains
02200930[c]	63	GA	41	0.404	14.5	-82.3	32.9	Southeastern Plains
02201000	64	GA	28	-0.521	110	-82.7	33.0	Southeastern Plains
02202000	65	GA	67	-0.442	1933	-82.4	33.1	Southeastern Plains
02202500[r]	66	GA	71	-0.548	2665	-82.2	32.9	Southeastern Plains
02202600[c]	67	GA	27	-0.219	227	-81.6	32.3	Southern Coastal Plain
02202800[c]	68	GA	26	-0.405	46.5	-82.3	32.7	Southeastern Plains
02203000	69	GA	69	-0.930	560	-82.1	32.4	Southeastern Plains
02203280[r]	70	GA	38	-0.289	838	-82.0	32.4	Southeastern Plains
02204135[c]	71	GA	30	-0.160	0.27	-84.2	33.6	Piedmont
02204500	72	GA	33	0.169	450	-84.2	33.6	Piedmont
02208450	73	GA	34	-0.029	180	-83.8	33.8	Piedmont
02212600	74	GA	42	0.031	72.5	-83.7	33.2	Piedmont
02213000[r]	75	GA	114	-0.417	2248	-84.0	33.5	Piedmont
02213050	76	GA	33	-0.167	31.3	-83.6	33.0	Piedmont
02214000[c]	77	GA	34	0.262	142	-84.0	32.8	Piedmont
02214280[c]	78	GA	28	-0.966	33.2	-83.4	32.7	Southeastern Plains

Table A (Continued)

USGS Hydrologic Unit Code	Site Index #	State	Years of Record	Sample at-site log skew	Drainage Area (mi^2)	Basin Centroid Location		Physigraphic Provience (underline indicates <70% of basin in region)
						Lat	Lon	
02214500	79	GA	34	-0.109	102	-83.9	32.5	Southeastern Plains
02215000	80	GA	86	-0.507	3764	-83.9	33.2	Piedmont
02215245[c]	81	GA	43	-1.45	1.26	-83.4	32.0	Southeastern Plains
02215500[r]	82	GA	99	-0.097	5229	-83.7	32.8	Piedmont
02216000[c]	83	GA	39	0.110	350	-83.1	32.3	Southeastern Plains
02217400[c]	84	GA	42	0.486	2.54	-83.7	34.1	Piedmont
02217500	85	GA	71	-0.554	392	-83.7	34.1	Piedmont
02217900	86	GA	33	0.362	289	-83.5	34.2	Piedmont
02218300[r]	87	GA	37	-0.384	942	-83.6	34.1	Piedmont
02218500[r]	88	GA	83	0.285	1076	-83.5	34.0	Piedmont
02219000	89	GA	34	-0.417	176	-83.7	33.9	Piedmont
02219500	90	GA	49	-0.127	444	-83.6	33.8	Piedmont
02220550	91	GA	26	-1.47	16.8	-83.0	33.4	Piedmont
02220900	92	GA	36	-0.527	266	-83.6	33.5	Piedmont
02221000	93	GA	25	-0.313	23.2	-83.7	33.4	Piedmont
02221525	94	GA	29	-0.127	190	-83.6	33.4	Piedmont
02223200	95	GA	28	-0.404	192	-83.3	32.9	Sand Hills
02223349[c]	96	GA	30	-0.332	0.428	-83.2	32.8	Southeastern Plains
02224000	97	GA	25	0.063	61.3	-83.2	32.5	Southeastern Plains
02224500	98	GA	58	-0.831	5102	-83.3	33.3	Piedmont
02225000[r]	99	GA	47	-0.341	11552	-83.4	33.0	Piedmont
02225200	100	GA	26	-0.066	64.1	-82.7	32.9	Southeastern Plains
02225250[c]	101	GA	29	0.066	224	-82.5	32.7	Southeastern Plains
02225330[c]	102	GA	41	0.060	9.83	-82.2	32.3	Southeastern Plains
02225500	103	GA	73	-0.845	1132	-82.5	32.5	Southeastern Plains
02226000[r]	104	GA	82	-0.040	13634	-83.3	32.9	Southeastern Plains
02226100	105	GA	42	-1.15	180	-81.9	31.5	Southern Coastal Plain
02226200[c]	106	GA	26	0.056	222	-83.0	31.5	Southeastern Plains
02226500	107	GA	70	-0.429	1251	-82.8	31.4	Southern Coastal Plain
02227000	108	GA	35	-0.519	140	-82.6	31.7	Southern Coastal Plain
02227200[c]	109	GA	31	-0.589	94.2	-82.6	31.6	Southern Coastal Plain
02227400[c]	110	GA	31	0.375	108	-82.5	31.8	Southern Coastal Plain
02227422	111	GA	31	-0.540	0.38	-82.3	31.4	Southern Coastal Plain
02227430	112	GA	30	-0.758	61.9	-82.1	31.7	Southern Coastal Plain
02227500	113	GA	56	-0.459	657	-82.3	31.7	Southern Coastal Plain
02227990[c]	114	GA	29	0.230	0.59	-81.8	31.2	Southern Coastal Plain
02228000[r]	115	GA	76	-0.279	2834	-82.5	31.5	Southern Coastal Plain
02314500	116	GA	73	-0.961	1129	-82.5	31.0	Southern Coastal Plain
02314600	117	GA	25	-1.11	93.7	-82.9	31.1	Southern Coastal Plain

Table A (Continued)

USGS Hydrologic Unit Code	Site Index #	State	Years of Record	Sample at-site log skew	Drainage Area (mi^2)	Basin Centroid Location		Physigraphic Provience (underline indicates <70% of basin in region)
						Lat	Lon	
02314700[r]	118	GA	25	-0.136	195	-82.9	31.0	Southern Coastal Plain
02315700[c]	119	GA	27	-0.534	117	-83.5	31.9	Southeastern Plains
02315900[c]	120	GA	26	-0.422	135	-83.7	31.8	Southeastern Plains
02316000	121	GA	43	-0.332	656	-83.5	31.7	Southeastern Plains
02316200	122	GA	28	-0.172	90.0	-83.2	31.7	Southeastern Plains
02317500	123	GA	79	-0.270	1343	-83.3	31.5	Southeastern Plains
02317700	124	GA	27	-0.640	125	-83.3	31.3	Southeastern Plains
02317710[c]	125	GA	28	-0.175	0.628	-83.3	31.2	Southeastern Plains
02317810[c]	126	GA	37	-0.700	0.158	-83.6	31.4	Southeastern Plains
02317900	127	GA	28	-0.377	45.5	-83.7	31.5	Southeastern Plains
02318000	128	GA	42	0.061	577	-83.7	31.4	Southeastern Plains
02318500	129	GA	36	-0.729	1494	-83.5	31.3	Southeastern Plains
02318700	130	GA	27	-0.052	272	-83.7	31.0	Southeastern Plains
02327200[c]	131	GA	27	-0.150	89.9	-83.9	31.3	Southeastern Plains
02327350	132	GA	42	-0.146	1.98	-84.0	31.0	Southeastern Plains
02327355[cr]	133	GA	26	-0.225	256	-83.9	31.2	Southeastern Plains
02327500	134	GA	51	0.384	559	-83.9	31.1	Southeastern Plains
02327700[c]	135	GA	27	-0.027	107	-84.1	31.0	Southeastern Plains
02327860[c]	136	GA	26	-0.159	1.66	-84.4	30.9	Southeastern Plains
02327900[r]	137	GA	27	-0.967	18.5	-84.3	30.9	Southeastern Plains
02328000	138	GA	36	-0.114	58.2	-84.3	30.9	Southeastern Plains
02330450	139	GA	26	0.242	44.8	-83.8	34.8	Blue Ridge
02331000	140	GA	60	-0.365	151	-83.7	34.7	Piedmont
02331500	141	GA	34	0.259	155	-83.5	34.7	Piedmont
02331600[r]	142	GA	67	-0.635	316	-83.6	34.7	Piedmont
02333500	143	GA	71	-0.037	151	-83.9	34.6	Piedmont
02335700	144	GA	46	-0.267	73.4	-84.2	34.2	Piedmont
02337000	145	GA	72	0.025	239	-84.7	33.8	Piedmont
02337400	146	GA	27	-0.496	47.0	-84.9	33.7	Piedmont
02337448[c]	147	GA	30	-1.12	0.31	-84.9	33.6	Piedmont
02338660	148	GA	28	-0.264	125	-84.9	33.3	Piedmont
02340250[c]	149	GA	29	0.211	201	-84.9	32.9	Piedmont
02340500	150	GA	30	0.392	61.8	-85.0	32.8	Piedmont
02341600[c]	151	GA	44	0.380	47.0	-84.5	32.5	Sand Hills
02341723	152	GA	27	1.49	31.4	-84.6	32.4	Sand Hills
02341800	153	GA	38	0.953	341	-84.7	32.5	Sand Hills
02341900	154	GA	28	0.869	53.4	-84.7	32.3	Southeastern Plains
02343200	155	GA	30	0.349	71.1	-84.7	32.0	Southeastern Plains
02343219[c]	156	GA	30	1.17	2.94	-84.9	32.1	Southeastern Plains

Table A (Continued)

USGS Hydrologic Unit Code	Site Index #	State	Years of Record	Sample at-site log skew	Drainage Area (mi^2)	Basin Centroid Location		Physigraphic Provience (underline indicates <70% of basin in region)
						Lat	Lon	
02343225[c]	157	GA	28	0.622	294	-84.8	32.0	Southeastern Plains
02343267[c]	158	GA	28	0.767	2.64	-85.0	31.4	Southeastern Plains
02344500	159	GA	70	-0.205	268	-84.4	33.4	Piedmont
02344700	160	GA	42	0.267	100	-84.6	33.4	Piedmont
02345000	161	GA	41	-0.706	963	-84.5	33.3	Piedmont
02346180[r]	162	GA	73	-0.454	1215	-84.5	33.2	Piedmont
02346217[c]	163	GA	32	-0.353	2.83	-84.6	32.8	Piedmont
02346500	164	GA	36	-0.312	188	-84.3	33.1	Piedmont
02347500[r]	165	GA	88	-0.318	1848	-84.5	33.1	Piedmont
02349000	166	GA	34	0.782	82.2	-84.4	32.5	Sand Hills
02349030[c]	167	GA	27	1.44	40.7	-84.4	32.4	Sand Hills
02349330[c]	168	GA	30	-0.836	0.39	-84.4	32.3	Sand Hills
02349350	169	GA	28	-0.004	150	-84.4	32.4	Sand Hills
02349605	170	GA	102	-0.207	2922	-84.4	32.9	Piedmont
02349695	171	GA	30	-0.663	0.69	-83.9	32.4	Southeastern Plains
02349900[c]	172	GA	56	-0.711	47.1	-83.8	32.2	Southeastern Plains
02350512[r]	173	GA	49	0.191	3929	-84.3	32.7	Piedmont
02350600	174	GA	51	0.143	197	-84.6	32.2	Sand Hills
02350685[c]	175	GA	29	0.310	0.30	-84.4	32.0	Southeastern Plains
02350900	176	GA	41	0.678	527	-84.5	32.1	Southeastern Plains
02351500[c]	177	GA	26	0.390	140	-84.4	32.2	Southeastern Plains
02351800[c]	178	GA	29	-0.413	46.7	-84.3	32.0	Southeastern Plains
02351890	179	GA	27	1.59	365	-84.3	32.1	Southeastern Plains
02352500[r]	180	GA	114	0.079	5282	-84.3	32.5	Southeastern Plains
02353000	181	GA	69	-0.069	5749	-84.3	32.4	Southeastern Plains
02353400	182	GA	49	0.669	182	-84.7	31.7	Southeastern Plains
02353500[r]	183	GA	69	0.163	627	-84.6	31.7	Southeastern Plains
02354500	184	GA	49	-0.559	318	-84.4	31.6	Southeastern Plains
02356000[r]	185	GA	98	-0.079	7528	-84.3	32.2	Southeastern Plains
02356100	186	GA	25	-0.316	46.8	-84.8	31.5	Southeastern Plains
02357000	187	GA	65	-0.613	468	-84.8	31.3	Southeastern Plains
02379500	188	GA	48	0.050	134	-84.3	34.6	Blue Ridge
02380500[r]	189	GA	64	-0.118	236	-84.4	34.7	Blue Ridge
02381600	190	GA	41	0.003	8.93	-84.5	34.6	Blue Ridge
02382200	191	GA	42	0.228	120	-84.5	34.5	Blue Ridge
02384500	192	GA	49	0.079	252	-84.7	35.0	Ridge and Valley
02384600	193	GA	43	-0.028	3.78	-84.8	34.8	Ridge and Valley
02385800	194	GA	46	0.238	64.1	-84.7	34.8	Blue Ridge
02387000[r]	195	GA	69	-0.201	687	-84.8	34.9	Ridge and Valley

Table A (Continued)

USGS Hydrologic Unit Code	Site Index #	State	Years of Record	Sample at-site log skew	Drainage Area (mi^2)	Basin Centroid Location		Physigraphic Provience (underline indicates <70% of basin in region)
						Lat	Lon	
02388900[cr]	196	GA	29	0.634	69.5	-84.1	34.6	Piedmont
02389000	197	GA	39	-0.387	107	-84.1	34.5	Piedmont
02392000	198	GA	115	0.141	613	-84.2	34.4	Piedmont
02394400[e]	199	GA	27	-0.364	42.4	-84.9	33.9	Piedmont
02395120	200	GA	26	-1.05	32.4	-84.9	34.3	Ridge and Valley
02397410[r]	201	GA	27	-0.650	65.4	-85.3	33.9	Ridge and Valley
02397500	202	GA	36	-0.346	115	-85.3	34.0	Ridge and Valley
02398000	203	GA	69	-0.414	192	-85.3	34.6	Ridge and Valley
02411800	204	GA	26	-0.160	20.1	-85.1	33.8	Piedmont
02411900[r]	205	GA	27	0.665	236	-85.1	33.8	Piedmont
02411902[e]	206	GA	29	-0.334	0.12	-85.3	33.9	Piedmont
02413000	207	GA	29	-0.266	94.9	-85.0	33.7	Piedmont
02413200	208	GA	29	-0.728	219	-85.1	33.6	Piedmont
03545000	209	GA	60	-0.005	45.3	-83.7	34.8	Blue Ridge
03550500	210	GA	58	-0.053	74.9	-83.9	34.8	Blue Ridge
03560000	211	GA	31	0.473	70.6	-84.4	34.9	Blue Ridge
03567200	212	GA	27	0.320	73.6	-85.4	34.7	Ridge and Valley
03568933	213	GA	27	-0.303	146	-85.5	34.8	Ridge and Valley
02053200	214	NC	48	1.04	225	-77.2	36.4	Middle Atlantic Coastal Plain
02053500	215	NC	56	0.905	63.0	-77.1	36.3	Middle Atlantic Coastal Plain
02070500[r]	216	NC	56	0.500	242	-80.1	36.6	Piedmont
02071000[r]	217	NC	67	-0.387	1053	-80.2	36.5	Piedmont
02077200	218	NC	39	-0.277	46.0	-79.2	36.3	Piedmont
02080500[r]	219	NC	95	0.505	8384	-79.4	36.8	Piedmont
02081500	220	NC	67	-0.452	167	-78.7	36.3	Piedmont
02081747	221	NC	43	0.337	427	-78.6	36.3	Piedmont
02082000[r]	222	NC	42	0.342	701	-78.4	36.2	Piedmont
02082770	223	NC	43	0.653	166	-78.2	36.2	Piedmont
02082950	224	NC	47	1.33	177	-77.9	36.3	Piedmont
02083000[r]	225	NC	92	0.418	526	-78.0	36.3	Piedmont
02083500	226	NC	70	0.581	2183	-78.1	36.2	Piedmont
02084500	227	NC	30	-0.180	10.0	-77.0	35.6	Middle Atlantic Coastal Plain
02084540	228	NC	39	0.227	26.0	-76.9	35.3	Middle Atlantic Coastal Plain
02084557	229	NC	29	-0.166	23.0	-76.8	35.7	Middle Atlantic Coastal Plain
02085000	230	NC	54	0.176	66.0	-79.2	36.1	Piedmont
2085070[r]	231	NC	43	-0.758	141	-79.1	36.1	Piedmont
0208521324	232	NC	19	-0.512	78.0	-79.0	36.2	Piedmont
02085500	233	NC	81	-0.501	149	-79.0	36.3	Piedmont
02086000	234	NC	47	0.358	5.00	-78.9	36.2	Piedmont

Table A (Continued)

USGS Hydrologic Unit Code	Site Index #	State	Years of Record	Sample at-site log skew	Drainage Area (mi^2)	Basin Centroid Location		Physigraphic Provience (underline indicates <70% of basin in region)
						Lat	Lon	
2087000[r]	235	NC	53	0.452	535	-79.0	36.2	Piedmont
02087500	236	NC	79	0.245	1150	-78.8	36.0	Piedmont
02087570[r]	237	NC	58	-0.176	1206	-78.8	36.0	Piedmont
02088000	238	NC	67	0.423	84.0	-78.7	35.6	Piedmont
02088470[r]	239	NC	25	0.121	191	-78.3	35.8	Piedmont
02088500	240	NC	76	0.569	232	-78.3	35.7	Piedmont
02089000	241	NC	77	0.439	2399	-78.6	35.8	Piedmont
02089500[r]	242	NC	79	0.141	2692	-78.5	35.7	Piedmont
02091000	243	NC	52	1.73	80.0	-77.9	35.5	Southeastern Plains
02091500	244	NC	78	0.295	733	-78.0	35.7	Southeastern Plains
02091700	245	NC	31	0.392	93.0	-77.6	35.6	Southeastern Plains
02092000	246	NC	39	-0.155	182	-77.3	35.5	Middle Atlantic Coastal Plain
02092500	247	NC	55	0.755	168	-77.6	35.0	Middle Atlantic Coastal Plain
02093000	248	NC	43	0.731	94.0	-77.6	34.9	Middle Atlantic Coastal Plain
02093800	249	NC	51	-0.546	21.0	-80.0	36.1	Piedmont
02094000	250	NC	30	1.12	16.0	-79.9	36.1	Piedmont
02095000	251	NC	37	0.393	34.0	-79.8	36.0	Piedmont
02096500	252	NC	78	0.007	606	-79.7	36.2	Piedmont
02096960[r]	253	NC	32	-0.295	1275	-79.5	36.1	Piedmont
02098500	254	NC	42	0.448	32.0	-80.0	36.1	Piedmont
02099500	255	NC	75	0.016	125	-79.9	36.0	Piedmont
02100500	256	NC	84	0.087	349	-79.8	35.9	Piedmont
02101000	257	NC	32	1.45	137	-79.7	35.4	Piedmont
02101800	258	NC	36	-0.022	16.0	-79.4	35.7	Piedmont
02102000	259	NC	76	0.327	1434	-79.6	35.6	Piedmont
02102500[r]	260	NC	83	-0.020	3464	-79.4	35.8	Piedmont
02102908	261	NC	38	-0.191	8.00	-79.2	35.2	Sand Hills
02103500	262	NC	44	0.653	459	-79.2	35.2	Sand Hills
02104000[r]	263	NC	71	0.468	4395	-79.3	35.7	Piedmont
02105500[r]	264	NC	60	-0.445	4852	-79.3	35.6	Piedmont
02105900	265	NC	34	0.024	22.0	-78.1	34.2	Middle Atlantic Coastal Plain
02106500	266	NC	55	0.487	676	-78.3	35.0	Southeastern Plains
02107000	267	NC	35	0.087	379	-78.6	35.1	Southeastern Plains
02108000	268	NC	66	0.506	599	-77.9	35.0	Southeastern Plains
02108500	269	NC	27	0.183	69.0	-78.1	34.8	Middle Atlantic Coastal Plain
02109500	270	NC	67	0.131	680	-78.5	34.3	Middle Atlantic Coastal Plain
02111000	271	NC	66	0.571	29.0	-81.6	36.1	Blue Ridge
02111180	272	NC	41	-0.030	48.0	-81.5	36.2	Blue Ridge
02111500	273	NC	65	0.175	89.0	-81.3	36.3	Blue Ridge

Table A (Continued)

USGS Hydrologic Unit Code	Site Index #	State	Years of Record	Sample at-site log skew	Drainage Area (mi^2)	Basin Centroid Location		Physigraphic Provience (underline indicates <70% of basin in region)
						Lat	Lon	
02112000	274	NC	93	0.981	504	-81.4	36.1	Blue Ridge
02112120	275	NC	42	-0.286	128	-81.1	36.3	Blue Ridge
02112360	276	NC	42	-0.952	79.0	-80.9	36.4	Piedmont
02113000	277	NC	85	0.475	128	-80.8	36.5	Piedmont
02113850	278	NC	42	-0.168	231	-80.6	36.5	Piedmont
02114450	279	NC	46	-0.592	43.0	-80.4	36.3	Piedmont
02116500	280	NC	78	-0.478	2280	-80.8	36.2	Piedmont
02117500	281	NC	31	-0.738	101	-80.9	36.0	Piedmont
02118000[r]	282	NC	68	-0.083	306	-80.9	35.9	Piedmont
02118500	283	NC	55	-1.35	155	-80.9	36.1	Piedmont
02119000[r]	284	NC	37	0.792	569	-80.9	36.0	Piedmont
02120780	285	NC	27	0.678	118	-80.7	35.7	Piedmont
02121500	286	NC	33	-0.411	174	-80.1	35.9	Piedmont
02123500	287	NC	32	0.420	342	-80.0	35.7	Piedmont
02125000	288	NC	52	-0.253	56.0	-80.4	35.4	Piedmont
02126000	289	NC	77	-0.377	1372	-80.5	35.2	Piedmont
02127000	290	NC	36	0.491	110	-80.3	34.9	Piedmont
02128000	291	NC	51	-0.126	106	-79.8	35.5	Piedmont
02133500	292	NC	67	1.14	183	-79.6	35.2	Sand Hills
02134500	293	NC	77	-0.176	1228	-79.2	34.8	Southeastern Plains
02137727[r]	294	NC	26	-0.238	126	-82.2	35.6	Blue Ridge
02138000	295	NC	40	0.393	172	-82.2	35.7	Blue Ridge
02138500	296	NC	84	0.584	67.0	-81.9	36.0	Blue Ridge
02142000	297	NC	51	-0.473	28.0	-81.2	36.0	Blue Ridge
02142900	298	NC	41	0.465	16.0	-80.9	35.3	Piedmont
02143000	299	NC	70	-0.251	83.0	-81.6	35.7	Blue Ridge
02143040	300	NC	45	-0.451	26.0	-81.6	35.6	Blue Ridge
02143500	301	NC	55	0.271	69.0	-81.4	35.5	Piedmont
02144000	302	NC	53	0.079	32.0	-81.3	35.3	Piedmont
02145000	303	NC	53	-0.125	628	-81.4	35.5	Piedmont
02146900	304	NC	44	0.091	77.0	-80.7	35.0	Piedmont
02149000	305	NC	55	-1.06	79.0	-82.1	35.5	Blue Ridge
02151000	306	NC	72	0.053	220	-81.9	35.4	Piedmont
02151500	307	NC	80	-0.454	875	-82.1	35.4	Piedmont
02152100	308	NC	46	-0.406	61.0	-81.8	35.5	Blue Ridge
03161000	309	NC	79	1.16	205	-81.6	36.3	Blue Ridge
03162500	310	NC	39	1.71	277	-81.6	36.5	Blue Ridge
03439000	311	NC	73	0.119	68.0	-82.9	35.2	Blue Ridge
03439500[r]	312	NC	31	0.353	103	-82.9	35.2	Blue Ridge

Table A (Continued)

USGS Hydrologic Unit Code	Site Index #	State	Years of Record	Sample at-site log skew	Drainage Area (mi^2)	Basin Centroid Location		Physigraphic Provience (underline indicates <70% of basin in region)
						Lat	Lon	
03441000	313	NC	83	-0.102	40.0	-82.8	35.3	Blue Ridge
03443000	314	NC	86	0.785	296	-82.8	35.2	Blue Ridge
03444500	315	NC	31	0.811	10.0	-82.8	35.4	Blue Ridge
03446000	316	NC	74	0.202	67.0	-82.7	35.4	Blue Ridge
03448000[r]	317	NC	52	0.298	676	-82.6	35.3	Blue Ridge
03449000	318	NC	32	0.687	24.0	-82.3	35.7	Blue Ridge
03450000	319	NC	72	0.116	5.00	-82.4	35.7	Blue Ridge
03451500	320	NC	111	0.502	945	-82.6	35.4	Blue Ridge
03453000	321	NC	52	0.399	158	-82.5	35.8	Blue Ridge
03453500[r]	322	NC	64	-0.276	1332	-82.6	35.5	Blue Ridge
03454000	323	NC	39	0.161	126	-82.7	36.0	Blue Ridge
03455500	324	NC	52	0.323	28.0	-82.9	35.3	Blue Ridge
03456500	325	NC	52	0.023	52.0	-82.8	35.4	Blue Ridge
03459500	326	NC	79	-0.015	350	-82.9	35.5	Blue Ridge
03460000	327	NC	62	-0.287	49.0	-83.1	35.6	Blue Ridge
03463300	328	NC	49	0.605	43.0	-82.2	35.8	Blue Ridge
03464000	329	NC	38	0.758	157	-82.4	35.9	Blue Ridge
03464500	330	NC	30	0.673	608	-82.2	36.0	Blue Ridge
03479000	331	NC	67	0.493	92.0	-81.8	36.2	Blue Ridge
03500000	332	NC	62	0.581	140	-83.4	35.0	Blue Ridge
03500240	333	NC	45	0.472	57.0	-83.5	35.1	Blue Ridge
03503000	334	NC	61	-0.310	436	-83.4	35.1	Blue Ridge
03504000	335	NC	67	0.124	52.0	-83.6	35.1	Blue Ridge
03512000	336	NC	58	-0.397	184	-83.3	35.6	Blue Ridge
03513000	337	NC	105	-0.054	655	-83.2	35.4	Blue Ridge
03513500	338	NC	36	0.162	14.0	-83.5	35.5	Blue Ridge
03548500	339	NC	107	-0.439	406	-83.8	35.0	Blue Ridge
03550000	340	NC	96	0.071	104	-83.8	35.2	Blue Ridge
02110500[r]	341	SC	56	-0.184	1125	-78.6	34.2	Middle Atlantic Coastal Plain
02130900	342	SC	47	0.339	107	-80.2	34.6	Sand Hills
02132100	343	SC	28	0.581	19.8	-79.8	33.9	Middle Atlantic Coastal Plain
02132500	344	SC	65	0.199	528	-79.5	34.7	Southeastern Plains
02135300	345	SC	38	0.125	97.9	-80.4	34.2	Sand Hills
02135500	346	SC	36	-0.449	386	-80.3	34.1	Southeastern Plains
02136000	347	SC	79	0.144	1236	-80.2	33.9	Southeastern Plains
02147500	348	SC	50	0.047	196	-81.1	34.6	Piedmont
02153500[r]	349	SC	51	-0.184	1499	-81.9	35.4	Piedmont
02154500	350	SC	74	-0.370	115	-82.2	35.2	Piedmont
02156500	351	SC	68	-0.638	2779	-81.8	35.2	Piedmont

Table A (Continued)

USGS Hydrologic Unit Code	Site Index #	State	Years of Record	Sample at-site log skew	Drainage Area (mi^2)	Basin Centroid Location		Physigraphic Provience (underline indicates <70% of basin in region)
						Lat	Lon	
02157000	352	SC	38	-0.477	44.2	-82.1	35.0	Piedmont
02157500	353	SC	51	-0.722	68.1	-82.2	35.1	Piedmont
02158000[r]	354	SC	41	-0.087	161	-82.1	35.0	Piedmont
02158500[r]	355	SC	41	0.040	105	-82.3	35.0	Piedmont
02159000	356	SC	44	0.079	172	-82.2	34.9	Piedmont
02159500[r]	357	SC	28	0.418	347	-82.2	35.0	Piedmont
02160000	358	SC	65	-0.925	186	-81.8	34.8	Piedmont
02160105[r]	359	SC	32	0.360	755	-81.9	34.8	Piedmont
02160500	360	SC	64	0.035	306	-82.2	34.8	Piedmont
02160700[r]	361	SC	33	1.06	443	-82.1	34.7	Piedmont
02161500[r]	362	SC	58	0.636	4823	-81.8	35.0	Piedmont
02162010	363	SC	29	-1.27	49.0	-81.0	34.3	Piedmont
02162500	364	SC	62	-0.470	296	-82.6	35.0	Blue Ridge
02163000[r]	365	SC	62	-0.292	411	-82.5	35.0	Piedmont
02163500	366	SC	68	-0.461	581	-82.5	34.9	Piedmont
02165000	367	SC	63	-0.132	236	-82.3	34.7	Piedmont
02165200	368	SC	30	0.349	29.8	-82.2	34.6	Piedmont
02169960	369	SC	26	0.111	1.25	-80.4	33.5	Southeastern Plains
02172500	370	SC	49	-0.061	196	-81.6	33.7	Sand Hills
02173000	371	SC	73	0.259	734	-81.5	33.6	Sand Hills
02173500	372	SC	68	0.141	686	-81.2	33.7	Southeastern Plains
02174000[r]	373	SC	60	-0.194	1726	-81.3	33.6	Sand Hills
02176500	374	SC	55	-0.601	196	-81.2	32.9	Middle Atlantic Coastal Plain
02185200	375	SC	37	-0.460	72.0	-83.0	34.9	Piedmont
02186000	376	SC	27	-0.431	104	-82.7	34.9	Piedmont
02192500	377	SC	64	-0.925	215	-82.5	34.3	Piedmont
02196000	378	SC	62	-0.403	544	-82.1	33.9	Piedmont
02384900	379	TN	31	0.458	4.41	-84.8	35.1	Ridge and Valley
03455000[r]	380	TN	91	0.068	1853	-82.7	35.6	Blue Ridge
03461200	381	TN	29	-0.250	10.3	-83.2	35.7	Blue Ridge
03461500[r]	382	TN	84	0.014	667	-83.0	35.6	Blue Ridge
03465000	383	TN	39	-0.207	15.9	-82.3	36.2	Blue Ridge
03465500[r]	384	TN	85	0.363	805	-82.3	36.0	Blue Ridge
03466500	385	TN	39	0.250	1184	-82.4	36.0	Blue Ridge
03467000	386	TN	27	0.190	219	-82.8	36.3	Ridge and Valley
03467500[e]	387	TN	61	0.341	1682	-82.5	36.2	Blue Ridge
03469130[cr]	388	TN	29	1.27	109	-83.4	35.7	Blue Ridge
03469160[e]	389	TN	29	0.524	63.7	-83.4	35.8	Blue Ridge
03469500	390	TN	32	-0.341	76.6	-83.5	35.7	Blue Ridge

Table A (Continued)

USGS Hydrologic Unit Code	Site Index #	State	Years of Record	Sample at-site log skew	Drainage Area (mi^2)	Basin Centroid Location		Physigraphic Provience (underline indicates <70% of basin in region)
						Lat	Lon	
03470000[r]	391	TN	63	-0.079	352	-83.5	35.8	Blue Ridge
03477000	392	TN	53	-0.505	812	-81.8	36.7	Ridge and Valley
03482000[c]	393	TN	40	0.231	102	-81.8	36.4	Blue Ridge
03483000	394	TN	28	2.20	427	-81.9	36.3	Blue Ridge
03485500	395	TN	66	0.234	137	-82.1	36.2	Blue Ridge
03487550[c]	396	TN	42	0.463	36.5	-82.4	36.6	Ridge and Valley
03491000	397	TN	58	-0.204	48.3	-82.9	36.5	Ridge and Valley
03491200[c]	398	TN	31	0.648	1.87	-83.0	36.4	Ridge and Valley
03491500	399	TN	40	-0.533	3033	-82.2	36.5	Ridge and Valley
03497000	400	TN	88	0.236	8902	-82.6	35.9	Blue Ridge
03497300	401	TN	42	0.220	106	-83.6	35.6	Blue Ridge
03498500[r]	402	TN	55	0.079	269	-83.7	35.6	Blue Ridge
03498700[c]	403	TN	31	0.600	0.37	-83.8	35.9	Ridge and Valley
03518500	404	TN	61	-0.649	118	-84.2	35.3	Blue Ridge
03519500	405	TN	65	0.126	2443	-83.7	35.3	Blue Ridge
03519610[c]	406	TN	33	0.216	1.67	-84.0	35.7	Ridge and Valley
03519640[c]	407	TN	33	0.826	16.6	-84.1	35.7	Ridge and Valley
03520100[c]	408	TN	29	0.524	61.8	-84.5	35.6	Ridge and Valley
03528000	409	TN	86	0.169	1475	-82.4	36.9	Ridge and Valley
03528400	410	TN	36	-0.123	2.62	-83.9	36.4	Ridge and Valley
03532000	411	TN	73	0.367	688	-83.1	36.7	Ridge and Valley
03533000[r]	412	TN	71	-0.031	2914	-82.9	36.8	Ridge and Valley
03534000[c]	413	TN	52	-0.479	24.4	-84.2	36.2	Central Appalachians
03534500	414	TN	31	-0.484	9.38	-84.0	36.2	Ridge and Valley
03535000	415	TN	34	0.487	68.6	-83.8	36.2	Ridge and Valley
03535180	416	TN	40	0.218	3.33	-83.9	36.1	Ridge and Valley
03538200[c]	417	TN	32	-0.246	56.2	-84.3	36.1	Central Appalachians
03538225[r]	418	TN	29	-0.101	82.8	-84.3	36.1	Central Appalachians
03538250	419	TN	28	-0.059	19.3	-84.3	36.0	Ridge and Valley
03538500	420	TN	48	0.193	83.0	-84.6	36.2	Southwestern Appalachians
03538600[c]	421	TN	35	0.167	12.0	-85.1	35.9	Southwestern Appalachians
03539500	422	TN	28	-0.679	93.6	-85.0	35.9	Southwestern Appalachians
03539800	423	TN	33	-0.320	519	-84.9	36.0	Southwestern Appalachians
03540500[r]	424	TN	78	-0.019	765	-84.8	36.1	Southwestern Appalachians
03541500	425	TN	44	0.251	109	-84.8	35.9	Southwestern Appalachians
03542500[c]	426	TN	16	0.846	96.6	-84.9	35.7	Southwestern Appalachians
03543500	427	TN	60	0.038	117	-84.7	35.6	Ridge and Valley
03544500	428	TN	52	-0.310	50.5	-85.1	35.6	Southwestern Appalachians
03556000[r]	429	TN	37	-0.650	27.5	-84.3	35.1	Blue Ridge

Table A (Continued)

USGS Hydrologic Unit Code	Site Index #	State	Years of Record	Sample at-site log skew	Drainage Area (mi^2)	Basin Centroid Location		Physigraphic Provience (underline indicates <70% of basin in region)
						Lat	Lon	
03557000	430	TN	42	-0.912	1223	-84.1	35.0	Blue Ridge
03560500	431	TN	35	-0.369	5.15	-84.4	35.0	Blue Ridge
03561000	432	TN	36	0.614	13.2	-84.3	35.0	Blue Ridge
03565300	433	TN	30	0.225	31.6	-84.8	35.1	Ridge and Valley
03565500	434	TN	37	0.233	57.9	-84.6	35.4	Ridge and Valley
03566000	435	TN	60	-0.932	2299	-84.2	35.0	Blue Ridge
03566200	436	TN	31	1.42	9.64	-85.0	35.1	Ridge and Valley
03566420	437	TN	39	0.849	19.1	-85.0	35.0	Ridge and Valley
03567500	438	TN	64	-0.219	427	-85.2	34.9	Ridge and Valley
03568000[r]	439	TN	131	-0.632	21373	-83.4	35.8	Ridge and Valley
03570800	440	TN	28	-0.454	15.4	-85.4	35.4	Southwestern Appalachians
03571000	441	TN	78	-0.186	400	-85.2	35.5	Southwestern Appalachians
03571800[c]	442	TN	50	0.645	50.5	-85.8	35.2	Southwestern Appalachians
02043500	443	VA	37	0.167	23.5	-76.6	36.6	Middle Atlantic Coastal Plain
02044000	444	VA	49	0.160	38.7	-78.3	37.1	Piedmont
02044200	445	VA	39	0.275	0.37	-78.2	37.0	Piedmont
02044500	446	VA	56	0.314	317	-78.1	37.0	Piedmont
02045500[r]	447	VA	77	0.177	577	-77.9	37.0	Piedmont
02046000	448	VA	60	0.161	113	-77.7	37.1	Piedmont
02046500	449	VA	26	-1.05	4.99	-77.3	36.9	Southeastern Plains
02047000	450	VA	65	0.421	1441	-77.6	37.0	Piedmont
02049700	451	VA	27	0.331	8.43	-77.0	36.8	Middle Atlantic Coastal Plain
02050050	452	VA	40	0.414	2.72	-76.8	36.6	Middle Atlantic Coastal Plain
02051000	453	VA	58	0.043	56	-78.4	37.1	Piedmont
02051500	454	VA	79	0.451	552	-78.2	36.9	Piedmont
02051600	455	VA	38	0.246	30.8	-78.0	36.9	Piedmont
02052000[r]	456	VA	56	-0.010	744	-78.1	36.8	Piedmont
02052500	457	VA	43	0.114	68.7	-77.8	36.6	Piedmont
02054500	458	VA	63	-0.281	254	-80.3	37.2	Ridge and Valley
02055000[r]	459	VA	108	-0.308	384	-80.2	37.2	Ridge and Valley
02055100	460	VA	50	-0.568	11.7	-79.9	37.4	Ridge and Valley
02056650	461	VA	32	-0.075	55.8	-80.0	37.2	Blue Ridge
02056900	462	VA	30	-1.05	115	-80.0	37.1	Piedmont
02057000[r]	463	VA	40	0.034	208	-79.9	37.1	Piedmont
02057500	464	VA	38	0.169	1017	-80.0	37.2	Ridge and Valley
02058400	465	VA	43	-0.349	351	-79.8	36.9	Piedmont
02058500[r]	466	VA	33	0.663	382	-79.8	36.9	Piedmont
02059500	467	VA	79	0.088	188	-79.7	37.3	Piedmont
02060500[r]	468	VA	32	0.143	1782	-79.8	37.1	Piedmont

Table A (Continued)

USGS Hydrologic Unit Code	Site Index #	State	Years of Record	Sample at-site log skew	Drainage Area (mi^2)	Basin Centroid Location		Physigraphic Provience (underline indicates <70% of basin in region)
						Lat	Lon	
02061300[c]	469	VA	26	0.370	4.87	-79.5	37.3	Piedmont
02061500	470	VA	70	0.092	315	-79.4	37.4	Piedmont
02062500[r]	471	VA	39	0.532	2404	-79.7	37.2	Piedmont
02064000	472	VA	70	0.835	165	-79.0	37.2	Piedmont
02065300[c]	473	VA	29	0.050	2.2	-78.8	37.3	Piedmont
02065500	474	VA	60	0.179	97.6	-78.8	37.2	Piedmont
02066500	475	VA	25	0.069	135	-78.6	37.1	Piedmont
02069700	476	VA	44	0.750	85.5	-80.3	36.6	Piedmont
02070000	477	VA	78	0.613	108	-80.1	36.7	Piedmont
02075500	478	VA	55	-0.160	2587	-79.9	36.5	Piedmont
02076000[r]	479	VA	35	0.789	2762	-79.8	36.6	Piedmont
02076500	480	VA	48	-0.170	9.18	-79.4	37.0	Piedmont
02079640	481	VA	41	-0.330	53.5	-78.4	36.7	Piedmont
03164000[r]	482	VA	77	0.481	1141	-81.4	36.5	Blue Ridge
03165000	483	VA	62	0.169	39.4	-80.9	36.6	Blue Ridge
03165500	484	VA	63	0.945	1350	-81.3	36.5	Blue Ridge
03167000	485	VA	88	0.206	258	-81.1	37.0	Ridge and Valley
03168750	486	VA	50	0.299	4.70	-80.8	37.1	Ridge and Valley
03170000	487	VA	78	-0.202	309	-80.4	37.0	Blue Ridge
03173000	488	VA	69	-0.005	299	-80.9	37.2	Ridge and Valley
03175500	489	VA	77	-0.605	223	-81.2	37.2	Ridge and Valley

APPENDIX B

REDUNDANT SITE PAIRS

This appendix contains Table B1, which lists the 141 site-pairs identified by the redundant screening algorithm, Figure 4.3, Section 4.3.7. The table contains the USGS site number, the U.S. State of the gauge site, as well as the normalized distance ND and the drainage area ratio DAR. Those sites which are underlined are the sites that were removed from the study due to redundancy. This appendix also contains Table B2 which lists the redundant sites that were removed from the analysis. Table B3 lists the additional 59 sites removed due to censored annual peak flow values.

Table B1: Southeastern U.S. redundant site pairs. Underlined USGS Hydrologic Unit Code means site was removed from analysis to address redundancy.

USGS Hydrologic Unit Code	State	USGS Hydrologic Unit Code	State	Normalized Distance (ND)	Drainage Area Ratio (DAR)
02399000	AL	*02399200*	AL	0.43	1.59
02404000	AL	*02404400*	AL	0.40	1.73
02404000	AL	*02404500*	AL	0.42	1.80
02404400	AL	*02404500*	AL	0.03	1.04
02412000	AL	*02412500*	AL	0.48	1.77
02413300	AL	*02413500*	AL	0.29	1.46
02315000	FL	*02315500*	FL	0.09	1.16
02315000	FL	*02315550*	FL	0.15	1.26
02315500	FL	*02315550*	FL	0.06	1.09
02319000	FL	*02319500*	FL	0.45	3.25
02329500	FL	*02329600*	FL	0.16	1.29
02191300	GA	*02192000*	GA	0.44	1.88
02197830	GA	02198000	GA	0.38	1.37
02200400	GA	*02200500*	GA	0.26	4.22
02202000	GA	*02202500*	GA	0.33	1.38
02203000	GA	*02203280*	GA	0.19	1.49
02204500	GA	*02213000*	GA	0.49	4.99
02213000	GA	02215000	GA	0.43	1.67
02215000	GA	*02215500*	GA	0.36	1.39
02217500	GA	*02218300*	GA	0.39	2.40
02217900	GA	*02218300*	GA	0.33	3.25
02217900	GA	*02218500*	GA	0.45	3.72
02218300	GA	*02218500*	GA	0.12	1.14
02215000	GA	*02225000*	GA	0.36	3.07
02215500	GA	*02225000*	GA	0.20	2.21
02224500	GA	*02225000*	GA	0.27	2.26
02215000	GA	*02226000*	GA	0.49	3.62
02215500	GA	*02226000*	GA	0.28	2.61
02224500	GA	*02226000*	GA	0.32	2.67
02225000	GA	*02226000*	GA	0.11	1.18
02226500	GA	*02228000*	GA	0.39	2.27
02227500	GA	*02228000*	GA	0.50	4.31
02315000	FL	02314500	GA	0.21	1.86
02315500	FL	02314500	GA	0.29	2.15
02315550	FL	02314500	GA	0.36	2.34
02314600	GA	*02314700*	GA	0.49	2.08
02319000	FL	02318500	GA	0.19	1.44
02329000	FL	02327500	GA	0.49	2.05

Table B1 (Continued)

USGS Hydrologic Unit Code	State	USGS Hydrologic Unit Code	State	Normalized Distance (ND)	Drainage Area Ratio (DAR)
02327355	GA	02327500	GA	0.36	2.18
02327900	GA	02328000	GA	0.47	3.15
02331000	GA	*02331600*	GA	0.36	2.09
02331500	GA	*02331600*	GA	0.33	2.04
02345000	GA	*02346180*	GA	0.15	1.26
02345000	GA	*02347500*	GA	0.38	1.92
02346180	GA	*02347500*	GA	0.24	1.52
02347500	GA	02349605	GA	0.33	1.58
02349605	GA	*02350512*	GA	0.28	1.34
02349605	GA	*02352500*	GA	0.46	1.81
02350512	GA	*02352500*	GA	0.20	1.34
02350512	GA	02353000	GA	0.28	1.46
02352500	GA	02353000	GA	0.08	1.09
02353400	GA	*02353500*	GA	0.31	3.46
02350512	GA	*02356000*	GA	0.48	1.92
02352500	GA	*02356000*	GA	0.28	1.43
02353000	GA	*02356000*	GA	0.20	1.31
02379500	GA	*02380500*	GA	0.32	1.76
02384500	GA	*02387000*	GA	0.38	2.73
02388900	GA	02389000	GA	0.38	1.54
02397410	GA	02397500	GA	0.24	1.75
02398300	AL	02398000	GA	0.47	1.91
02412000	AL	*02411900*	GA	0.41	1.90
02413300	AL	02413200	GA	0.35	1.85
02070500	NC	*02071000*	NC	0.50	4.35
02081747	NC	*02082000*	NC	0.41	1.64
02082950	NC	*02083000*	NC	0.30	2.97
02083000	NC	02083500	NC	0.29	4.15
02085000	NC	*02085070*	NC	0.46	2.14
02085500	NC	*02087000*	NC	0.49	3.59
02087500	NC	*02087570*	NC	0.05	1.05
02088470	NC	02088500	NC	0.19	1.21
02087570	NC	02089000	NC	0.48	1.99
02089000	NC	*02089500*	NC	0.12	1.12
02096500	NC	*02096960*	NC	0.38	2.10
02096960	NC	*02102500*	NC	0.41	2.72
02102000	NC	*02102500*	NC	0.31	2.42
02102000	NC	*02104000*	NC	0.27	3.06

Table B1 (Continued)

USGS Hydrologic Unit Code	State	USGS Hydrologic Unit Code	State	Normalized Distance (ND)	Drainage Area Ratio (DAR)
02102500	NC	*02104000*	NC	0.15	1.27
02102000	NC	*02105500*	NC	0.28	3.38
02102500	NC	*02105500*	NC	0.22	1.40
02104000	NC	*02105500*	NC	0.07	1.10
02117500	NC	*02118000*	NC	0.27	3.03
02118000	NC	*02119000*	NC	0.20	1.86
02118500	NC	*02119000*	NC	0.46	3.67
02137727	NC	02138000	NC	0.17	1.37
03439000	NC	*03439500*	NC	0.34	1.51
03439500	NC	03443000	NC	0.44	2.87
03443000	NC	*03448000*	NC	0.50	2.28
03448000	NC	03451500	NC	0.19	1.40
03448000	NC	*03453500*	NC	0.39	1.97
03451500	NC	*03453500*	NC	0.21	1.41
02109500	NC	*02110500*	SC	0.22	1.65
02151500	NC	*02153500*	SC	0.33	1.71
02153500	SC	02156500	SC	0.29	1.85
02157000	SC	*02158000*	SC	0.22	3.64
02157500	SC	*02158500*	SC	0.46	1.54
02158000	SC	02159000	SC	0.40	1.07
02158000	SC	*02159500*	SC	0.18	2.16
02159000	SC	*02159500*	SC	0.16	2.02
02160000	SC	*02160105*	SC	0.42	4.07
02160105	SC	*02160700*	SC	0.45	1.71
02160500	SC	*02160700*	SC	0.38	1.45
02156500	SC	*02161500*	SC	0.26	1.74
02162500	SC	*02163000*	SC	0.27	1.39
02163000	SC	02163500	SC	0.39	1.41
02173000	SC	*02174000*	SC	0.36	2.35
02173500	SC	*02174000*	SC	0.26	2.52
03451500	NC	*03455000*	TN	0.36	1.96
03453500	NC	*03455000*	TN	0.21	1.39
03459500	NC	*03461500*	TN	0.39	1.90
03464500	NC	*03465500*	TN	0.12	1.32
03464500	NC	*03466500*	TN	0.38	1.95
03465500	TN	*03466500*	TN	0.27	1.47
03466500	TN	03467500	TN	0.35	1.42
03469130	TN	*03470000*	TN	0.48	3.24

Table B1 (Continued)

USGS Hydrologic Unit Code	State	USGS Hydrologic Unit Code	State	Normalized Distance (ND)	Drainage Area Ratio (DAR)
03469500	TN	*03470000*	TN	0.39	4.60
03455000	TN	03497000	TN	0.38	4.80
03497300	TN	*03498500*	TN	0.44	2.54
03532000	TN	*03533000*	TN	0.43	4.24
03538200	TN	*03538225*	TN	0.30	1.47
03539800	TN	*03540500*	TN	0.38	1.47
03557000	TN	*03566000*	TN	0.23	1.88
03497000	TN	*03568000*	TN	0.38	2.40
02044500	VA	*02045500*	VA	0.49	1.82
02051500	VA	*02052000*	VA	0.24	1.35
02054500	VA	*02055000*	VA	0.25	1.51
02056900	VA	*02057000*	VA	0.34	1.81
02055000	VA	02057500	VA	0.50	2.65
02057000	VA	02057500	VA	0.40	4.89
02058400	VA	*02058500*	VA	0.07	1.09
02057500	VA	*02060500*	VA	0.24	1.75
02080500	NC	*02062500*	VA	0.47	3.49
02057500	VA	*02062500*	VA	0.40	2.36
02060500	VA	*02062500*	VA	0.17	1.35
02070500	NC	02070000	VA	0.33	2.24
02071000	NC	02075500	VA	0.43	2.46
02080500	NC	02075500	VA	0.49	3.24
02071000	NC	*02076000*	VA	0.49	2.62
02080500	NC	*02076000*	VA	0.44	3.04
02075500	VA	*02076000*	VA	0.06	1.07
03162500	NC	*03164000*	VA	0.49	4.12
03164000	VA	03165500	VA	0.13	1.18

Table B2: Southeastern U.S. redundant sites removed from regional skew regression analysis (92 sites)

USGS Hydrologic Unit Code	State	USGS Hydrologic Unit Code	State	USGS Hydrologic Unit Code	State
02398300	AL	02347500	GA	02153500	SC
02399200	AL	02350512	GA	02158000	SC
02404400	AL	02352500	GA	02158500	SC
02404500	AL	02353500	GA	02159500	SC
02412500	AL	02356000	GA	02160105	SC
02413300	AL	02380500	GA	02160700	SC
02413500	AL	02387000	GA	02161500	SC
02315000	FL	02388900	GA	02163000	SC
02315500	FL	02397410	GA	02174000	SC
02315550	FL	02411900	GA	03455000	TN
02319000	FL	02070500	NC	03461500	TN
02319500	FL	02071000	NC	03465500	TN
02329000	FL	02080500	NC	03466500	TN
02329600	FL	02082000	NC	03469130	TN
02192000	GA	02083000	NC	03470000	TN
02197830	GA	02085070	NC	03498500	TN
02200500	GA	02087000	NC	03533000	TN
02202500	GA	02087570	NC	03538225	TN
02203280	GA	02088470	NC	03540500	TN
02213000	GA	02089500	NC	03566000	TN
02215500	GA	02096960	NC	03568000	TN
02218300	GA	02102500	NC	02045500	VA
02218500	GA	02104000	NC	02052000	VA
02225000	GA	02105500	NC	02055000	VA
02226000	GA	02118000	NC	02057000	VA
02228000	GA	02119000	NC	02058500	VA
02314700	GA	02137727	NC	02060500	VA
02327355	GA	03439500	NC	02062500	VA
02327900	GA	03448000	NC	02076000	VA
02331600	GA	03453500	NC	03164000	VA
02346180	GA	02110500	SC		

Table B3: Southeastern U.S. sites removed from regional skew regression analysis due to censored values (59 sites)

USGS Hydrologic Unit Code	State	USGS Hydrologic Unit Code	State
02329600	FL	02343225	GA
02191930	GA	02343267	GA
02200930	GA	02346217	GA
02202600	GA	02349030	GA
02202800	GA	02349330	GA
02204135	GA	02349900	GA
02214000	GA	02350685	GA
02214280	GA	02351500	GA
02215245	GA	02351800	GA
02216000	GA	02388900	GA
02217400	GA	02394400	GA
02223349	GA	02411902	GA
02225250	GA	03467500	TN
02225330	GA	03469130	TN
02226200	GA	03469160	TN
02227200	GA	03482000	TN
02227400	GA	03487550	TN
02227990	GA	03491200	TN
02315700	GA	03498700	TN
02315900	GA	03519610	TN
02317710	GA	03519640	TN
02317810	GA	03520100	TN
02327200	GA	03534000	TN
02327355	GA	03538200	TN
02327700	GA	03538600	TN
02327860	GA	03542500	TN
02337448	GA	03571800	TN
02340250	GA	02061300	VA
02341600	GA	02065300	VA
02343219	GA		

APPENDIX C

SENSITIVITY ANALYSIS ON REDUNDANT SITES

After completing the B-GLS regional skew regression of the Southeastern U.S., the topic of redundant sites was revisited. Section 4.3 considered the impact of nested watersheds on the B-GLS framework, as well as developed criteria for identifying these redundant gage sites. After the redundant pairs of gauge sites were identified based on ND and DAR thresholds, a decision was then made concerning which site would stay in the analysis and which site would be removed to eliminate redundant hydrologic experiences. Based on the procedure outlined in Section 4.3.7 and specifically in Figure 4.7, there is a preference to retain those sites with smaller drainage areas and longer records. This is due to the fact that the regional skew model being developed will most often be used to determine skew at small ungauged sites. Thus, there was a preference to keep in the analysis those sites which are most similar to the small, ungauged sites the regional skew will be employed to in the future.

This section changes the screening preferences for redundant sites. Previously, the preference for determining which site to retain from a redundant pair was to keep the site with the small drainage area and long record. Here, this preference is changed. Instead we remove those sites with small drainage areas and long record sites. Thus, using this modified criteria, the preference is now to retain those sites with large drainage areas and short record lengths. This was done to check the stability of the B-GLS results generated in Section 4.5 depending on which sites are included in the regional analysis.

The same thresholds used in Section 4.3.7 for both normalized distance ND and drainage area ratio DAR (0.5 and 5, respectively) were applied in this case. Figure C1 below outlines the modified algorithm used to screen sites

If ND< T_{ND} **&** DAR < T_{DAR}
IF (DA_{small} has >= 30 yrs of data), **THEN** (remove DA_{small})
ELSE IF [(DA_{small} <30 yrs of data) **&** (record length of DA_{small} + 5)] ≥ (record length of DA_{large}), **THEN** (remove DA_{small})
ELSE (remove DA_{large})

Figure C1: Modified screening algorithm for redundant sites

where T_{ND} is the normalized distance threshold, and T_{DAR} is the drainage area ratio threshold, DA_{small} is the site with the smaller drainage area, and DA_{large} is the site with the larger drainage area. The modified screening criteria in Figure C1 first identifies troublesome pairs: pairs of sites whose ND is less than T_{ND} and whose DAR is less than T_{DAR}. After identifying the troublesome pairs, the algorithm then runs through those troublesome pairs in index number order (see Section 4.2.1 for description of index number) to resolve each redundant site conflict by recommending the removal of one of the two sites.

Using the modified algorithm in Figure C1, the same set of redundant site pairs was identified as in Section 4.3.7, this is due to the fact that the thresholds for ND and DAR were the same. However, from these redundant pairs, a new set of sites were recommended to be removed from the study. According to the initial algorithm used in Section 4.3.7, 92 sites were removed from the study due to redundancy. Using the modified algorithm, 78 sites were removed from the study due to redundancy. A list of those 78 sites can be found in Table C1. Before the B-GLS regional skew regression could be run on the new data set, the 59 gage sites with censored data also

had to be removed. Thus, in total 137 sites were removed from the initial 489 sites, leaving 352 sites in the regional analysis. In Section 4.3.6, 147 sites were removed.

Table C1: Southeastern U.S. redundant sites removed (77 sites) from regional skew regression analysis based on the modified screening algorithm in Figure C1.

USGS Hydrologic Unit Code	State	USGS Hydrologic Unit Code	State	USGS Hydrologic Unit Code	State	USGS Hydrologic Unit Code	State
2404000	AL	2346180	GA	2096960	NC	2173000	SC
2315500	FL	2347500	GA	2102000	NC	2173500	SC
2319000	FL	2349605	GA	2102500	NC	3455000	TN
2191300	GA	2350512	GA	2109500	NC	3465500	TN
2202000	GA	2352500	GA	2117500	NC	3466500	TN
2203000	GA	2353000	GA	2151500	NC	3469500	TN
2204500	GA	2353400	GA	3439500	NC	3497000	TN
2213000	GA	2350900	GA	3443000	NC	3497300	TN
2215000	GA	2379500	GA	3448000	NC	3532000	TN
2215500	GA	2384500	GA	3451500	NC	3539800	TN
2217500	GA	2398000	GA	3453500	NC	2044500	VA
2217900	GA	2070500	NC	3459500	NC	2051500	VA
2224500	GA	2071000	NC	2153500	SC	2054500	VA
2225000	GA	2082950	NC	2156500	SC	2055000	VA
2226500	GA	2083000	NC	2157500	SC	2057000	VA
2227500	GA	2085500	NC	2158000	SC	2057500	VA
2314500	GA	2087500	NC	2160700	SC	2062500	VA
2327500	GA	2087570	NC	2162500	SC	2075500	VA
2331000	GA	2089000	NC	2163000	SC	3164000	VA
2345000	GA	2096500	NC				

After identifying and screening redundant gauge sites based on the modified criteria, in which there is a preference for those sites with large drainage areas and short records, B-GLS regression was employed to identify the best regional skewness model. The same cross correlation model (developed in Section 4.4.3) used in Section 4.5, is used here as well. The results from the B-GLS skew regression are shown below. Table C2 contains twenty-two single parameter models plus a constant model. A B-GLS regression was performed on all of the available explanatory variables. Each model contains a constant term and one explanatory variable.

Table C2: Single parameter B-GLS skew regression models for the Southeastern U.S. data set (352 sites), generated using the modified redundant site algorithm. Bayesian standard deviations and plausibility values, as percentages, are presented in ().

Model	Constant	Physiographic Province/ Basin Parameter	σ_δ^2	Avgerage Sampling Variance	AVP_{new}	R_δ^2
Constant	-0.012 (0.063)	-	0.141 (0.021)	0.0039	0.145	0.0%
Blue Ridge	0.002 (0.063)	0.002 (0.001) (4.0%)	0.139 (0.021)	0.0058	0.144	1.6%
Central Appalacians	-0.027 (0.063)	-0.002 (0.001) (2.7%)	0.139 (0.021)	0.0056	0.144	1.4%
Middle Atlanic Coastal Plain	-0.014 (0.063)	0.005 (0.002) (0.42%)	0.136 (0.021)	0.0050	0.141	3.1%
Piedmont	-0.020 (0.063)	0.001 (0.001) (16%)	0.141 (0.021)	0.0056	0.146	0.1%
Ridge and Valley	-0.013 (0.063)	0.0001 (0.002) (95%)	0.141 (0.021)	0.0051	0.147	-0.5%
Sand Hils	-0.013 (0.063)	0.001 (0.007) (89%)	0.142 (0.021)	0.0048	0.146	-0.6%
Southeastern Plains	-0.011 (0.063)	-0.002 (0.001) (28%)	0.140 (0.021)	0.0053	0.145	0.5%
Southern Coastal Plain	-0.004 (0.063)	-0.006 (0.002) (0.71%)	0.141 (0.021)	0.0059	0.146	0.1%
Southwestern Appaliacians	-0.009 (0.063)	-0.002 (0.002) (27%)	0.141 (0.021)	0.0051	0.146	-0.1%
Drainage Area (mi^2)	-0.005 (0.064)	-0.011 (0.018) (54%)	0.141 (0.021)	0.0050	0.146	-0.2%

Table C2 (Continued):

Model	Constant	Physiographic Province/ Basin Parameter	σ_δ^2	Avgerage Sampling Variance	AVP_{new}	R_δ^2
Main Channel Slope (ft/mi)	-0.011 (0.063)	0.005 (0.026) (83%)	0.141 (0.021)	0.0052	0.146	-0.3%
Average basin slope (%)	-0.008 (0.063)	0.003 (0.004) (43%)	0.141 (0.021)	0.0061	0.148	-0.5%
Main Channel Length (mi)	-0.006 (0.064)	-0.016 (0.029) (58%)	0.141 (0.021)	0.0050	0.146	-0.2%
Basin perimeter length (mi)	-0.007 (0.064)	-0.016 (0.031) (62%)	0.141 (0.021)	0.0049	0.146	-0.2%
Basin shape factor	-0.015 (0.063)	0.003 (0.006) (59%)	0.142 (0.021)	0.0048	0.146	-0.6%
Avg basin elev (ft, NAVD88)	-0.003 (0.063)	0.035 (0.035) (32%)	0.141 (0.021)	0.0059	0.147	0.0%
Max basin elev (ft, NAVD88)	-0.005 (0.063)	0.035 (0.033) (30%)	0.141 (0.021)	0.0059	0.147	-0.1%
Avg ann. Precip. in basin (in)	-0.010 (0.063)	0.004 (0.006) (58%)	0.140 (0.021)	0.0059	0.146	0.2%
% basin impervious surfaces	-0.016 (0.063)	-0.006 (0.007) (37%)	0.141 (0.021)	0.0050	0.146	-0.2%
% basin occupied by forests	-0.012 (0.063)	0.0002 (0.002) (89%)	0.141 (0.021)	0.0053	0.147	-0.5%
Avg soil drainage index	0.001 (0.063)	-0.104 (0.058) (7.4%)	0.141 (0.021)	0.0059	0.146	0.1%
Avg hydrologic soil index	0.001 (0.063)	-0.177 (0.096) (6.4%)	0.140 (0.021)	0.0055	0.146	0.2%

As shown in Table C2, none of the single parameter models significantly improved the model fit as compared to the constant model. In particular, all of the Pseudo R_δ^2 values were less than 4%. In comparing Table C2 to the single parameter models in Table 4.10 from Section 4.5, it is clear that there are only small differences resulting from using the original versus the modified site selection criteria. Focusing on the constant model, σ_δ^2 and AVP_{new} are the same in both cases. While there is only a slight increase of 0.0001 in the average sampling variance from the original B-GLS results to the B-GLS results obtained using the modified screening algorithm. The regression constant did decrease from the original B-GLS result of -0.019 to the modified B-GLS result of -0.012, however this is small difference. The standard deviation of the regression coefficient in both cases in 0.063, thus the difference in the regression constant is only about half a standard deviation.

As shown in Table C3, none of the multi-parameter models significantly improved the model fit as compared to the constant model. The Pseudo R_δ^2 values are still under 10% indicating that the use of several explanatory variables does not result in a major improvement in the fit as compared to the constant model. Alternatively, the addition of explanatory variables to the model does increase the complexity of the model. Thus, the constant model is again chosen as the best regional skew model.

Table C3: Multi-parameter B-GLS skew regression models for the Southeastern U.S. data set (342 sites). Bayesian standard deviations and plausibility values, as percentages, are presented in parenthesis.

Model	Constant	BR	SH	σ_δ^2	Avgerage Sampling Variance	AVP_{new}	R_δ^2
Constant	-0.012 (0.063)	-	-	0.141 (0.021)	0.0039	0.145	0.0%
BR	0.002 (0.063)	0.002 (0.001) (4.0%)	-	0.139 (0.021)	0.0058	0.144	1.6%
SH	-0.014 (0.063)	-	0.005 (0.002) (0.4%)	0.136 (0.021)	0.0050	0.141	3.1%
H	0.002 (0.063)	0.003 (0.001) (2.7%)	0.005 (0.002) (0.3%)	0.129 (0.020)	0.0070	0.136	6.9%

Thus, the B-GLS regional skew regression is stable and does not depend on which redundant site screening criteria, or more specifically, whether the redundant site screening criteria has a preference to retain those sites with small drainage areas and long records versus those sites with large drainage areas and short records, in this case.

REFERENCES

Gotvald, A.J., Feaster, T.D., and Weaver, J.C., (2009), Magnitude and frequency of rural floods in the southeastern United States, 2006—Volume 1, Georgia: U.S. Geological Survey Scientific Investigations Report 2009–5043, 120 p.

Feaster, T.D., Gotvald, A.J., and Weaver, J.C., (2009), Magnitude and frequency of rural floods in the southeastern United States, 2006—Volume *X*, South Carolina: U.S. Geological Survey Scientific Investigations Report (in progress).

Weaver, J.C., Feaster, T.D., and Gotvald, A.J., (2009), Magnitude and frequency of rural floods in the southeastern United States, 2006—Volume *X*, North Carolina: U.S. Geological Survey Scientific Investigations Report (in progress).

Gruber, A.M., D.S. Reis Jr., and J. R. Stedinger (2007), Models of regional skew based on Bayesian GLS regression, *World Environmental & Water Resources Conference-Restoring out Natural Habitat*, edited by K.C. Kabbes, Tampa, Florida May 15-18, Paper 40927-3285.

Griffis, V. W., J. R. Stedinger, and T. A. Cohn (2004), Log Pearson type 3 quantile estimators with regional skew information and low outlier adjustments, Water Resour. Res., 40, W07503, doi:10.1029/2003WR002697.

Hardison, C. H. (1971), Prediction error of regression estimates of streamflow characteristics at ungaged sites: U.S. Geological Survey Profession Paper, 750-C, pp. 228-236.

Interagency Advisory Committee on Water Data (1982), Guidelines for determining flood flow frequency, *Bulletin #17B*, U.S. Department of the Interior, U.S. Geological Survey, Office of Water Data Coordination, Reston, Virginia.

Kendall, M.G. and A. Stuart (1961), *The Advanced Theory of Statistics*, Vol 2, Hafner Publishing Company, New York.

Kenney, J. F. and E. S. Keeping (1951), *Mathematics of Statistics Part Two*, 2nd ed., pp. 217-221, D. Van Nostrand Company Inc, Princeton, NJ.

Maidment, D.R. (Ed.) (1993), *Handbook of Hydrology*, pp. 17.47-17.48, McGraw-Hill, Inc, New York.

Martins, E. S., and J.R. Stedinger (2002), Cross correlations among estimators of shape, *Water Resour. Res., 38*(11), 1252, doi:10.1029/2002WR001589.

Reis Jr., D.S., (2005). Flood Frequency Analysis Employing Bayesian Regional Regression and Imperfect Historical Information. Ph.D. Dissertation, Cornell University.

Reis, D. S., Jr., J. R. Stedinger, and E. S. Martins (2005), Bayesian generalized least squares regression with application to log Pearson type 3 regional skew estimation, *Water Resour. Res.*, *41*, W10419, doi:10.1029/2004WR003445.

Stedinger, J. R., and G. D. Tasker (1986b), Regional hydrologic analysis, 2: Model-error estimators, estimation o f sigma and log-Pearson type 3 distributions, *Water Res. Resear.*, *22*(10), pp. 1487-1499.

Stedinger, J.R., R.M. Vogel, and E. Foufoula-Georgiou, (1993), Frequency Analysis of Extreme Events, in Handbook of Hydrology, chao. 18, pp. 18.2 - 18.66, McGraw-Hill, New York.

Tasker, G. D., and J. R. Stedinger (1989), An Operational GLS Model for Hydrologic Regression, *Journal of Hydrology*, 111, pp. 361-375.

CHAPTER 5

CONCLUSIONS

5.1 Regional Hydrologic Regression Analysis

The research presented in this thesis develops new statistical techniques for estimating regional skewness coefficients to improve flood frequency analysis in the United States. Flood frequency guidelines for the United States, specified in *Bulletin 17B*, recommend fitting the log-Pearson Type III (LP3) distribution to the series of annual flood maxima, in which the third moment of the distribution, the skewness coefficient, is combined with a regional skewness coefficient to improve its precision. The research presented here extends the quasi-analytic Bayesian analysis of the Generalized Least Squares (GLS) regional hydrologic regression framework introduced by Reis *et al*. [2005] to more accurately and precisely estimate regional skewness coefficients. Specifically, formulas derived within a Bayesian regression framework for the computation of estimators, standard errors, and diagnostic statistics are provided by Reis [2005] and Reis *et al.* [2005]. Diagnostic statistics used as criteria for model selection include the Average Variance of Prediction (AVP), as well as the Bayesian Plausibility Value ψ. The Bayesian Plausibility Value originally developed by Reis *et al.* [2005] and later expanded on by Gruber *et al.* [2007] takes the place of the traditional p-value used in classical statistics; it describes whether zero is a plausible value for each β-parameter in a regression analysis given the prior distribution and the data. Regression diagnostic statistics for Weighted Least Squares (WLS) and GLS analyses include pseudo Analysis of Variance (ANOVA) tables, a pseudo adjusted R_{δ}^{2}, Error Variance Ratio (EVR) and Misrepresentation of the Beta Variance (MBV), leverage and influence, and σ-influence. EVR indicates if WLS or

GLS analysis is likely to be needed, or if an Ordinary Least Squares (OLS) analysis will suffice. Similarly, MBV indicates if a GLS analysis is needed, or if a WLS analysis would suffice. The R_{δ}^{2} statistic describes how well the model explains the variability in the true dependent variable, while the pseudo ANOVA table describes how much of the variation in the observations can be attributed to the model, and how much to the model error and sampling error. Finally, the leverage and influence metrics identify and consider the impact of any unusual observations on the models. Those metrics, as well as the newly developed σ-influence introduced by Gruber *et al*. [2007], allow for a comprehensive examination of the models developed from the B-GLS regression framework.

5.2 United States Flood Flow Frequency Procedures and Regional Skew

Currently, *Bulletin 17B* allows for regional skew values to be obtained from the skew map included with the Bulletin. As it is over 30 years old, the regional skew values from the *Bulletin 17B* skew map do not reflect annual maximum data acquired since 1976. This increase in available data, along with advances in computing power to support the Bayesian GLS regional hydrologic regression framework, allow for a much more precise estimate of the regional skewness coefficient for use in flood frequency analysis.

The recommended technique to perform flood frequency analyses, as described in Bulletin 17B, is to fit a log-Pearson Type III (LP3) distribution to the series of annual maximums. The third moment of the LP3 distribution is the skewness coefficient, which is very sensitive to extreme events, such as large floods. Thus, an accurate estimate of the skewness coefficient is important in flood frequency analysis because the majority of the interest is focused on the large flood events. However, short record lengths at gauged sites make a regional estimate of skew extremely

valuable in determining flood frequency estimates. Thus, the research documented here focuses on advancing a procedure to develop regional skewness estimators for flood frequency analysis using a Bayesian Generalized Least Squares (B-GLS) regression framework.

Two examples of regionalization of the log-space skew illustrate use of the methodology and compare the results obtained from OLS, WLS, and GLS analyses using both Bayesian and method of moments estimation techniques. The OLS analysis provides misleading results because it does not make a distinction between the variance due to the model error and the variance due to time sampling error. GLS is the best framework because the cross-correlation of the skews, which is neglected by WLS, proves to be important. These examples demonstrate that the model error variance for regional skew models is on the order of 0.15 or less. Leverage, influence and σ-influence statistics are very useful in identifying stations that actually did have a significant impact on the analysis.

The regional regression framework developed by Reis *et al*. [2005], along with the regression diagnostic statistics discussed in Chapter 2, are used to develop a regional skewness estimator for the Southeastern United States, nominally Georgia, North Carolina, and South Carolina. Criteria are adopted to identify redundant watersheds, those pairs of basins whose drainage areas are nested and of similar size, and thus cannot be considered independent hydrologic experiences. As discussed in Chapter 4, Normalized Distance (ND) is used to determine the likelihood that two drainage basins are nested, while the Drainage Area Ratio (DAR) is used to determine if two nested basins are sufficiently similar in size that they are essentially or are at least in large part the same watershed for the purposes of developing a regional hydrologic model. Although there are times when the combination of normalized distance and drainage area ratio will incorrectly identify distinct basins as redundant, it

appears that ND < 0.5 combined with a DAR threshold can successfully be used as a screening metric to identify redundant basins which represent the same hydrologic experience.

The Southeastern U.S. regional skewness study also provided an improved model for the estimation of cross-correlations of annual peaks using the Fisher Z transformation and an exponential transformation of the distance between basin centroids. In order to create a model with normal errors to describe the cross-correlations of annual peaks, the cross-correlation needs to be mapped into the whole real space [-∞,+∞] to match the use of an unbounded normal-error model. The Fisher Z transformation maps the sample correlation that is restricted on [-1,+1] to [-∞,+∞]. In Reis *et al.*[2005] and Tasker and Stedinger [1989], the inter-site correlation coefficient between concurrent flows $\rho(d_{ij})$ is modeled solely as a function of the distances between two site gauges. The gauges for each basin are located at the outlet of the basin, while the centroid is the geographical center of each basin. Thus, using the distance between basin centroids presents a better representation of the proximity and similarities of any basin pair. After redundant sites have been removed, the results of the Southeastern U.S. cross-correlation models show that the distance between basin centroids and the exponential transformation provide a great model fit ($R_{\delta}^2 = 83\%$).

Based upon a B-GLS analysis of the selected 342 stations, a constant generalized regional skew model is selected for the Southeastern U.S. Region described by the equation: $\hat{\gamma} = -0.019$, with MSE = 0.14. More complicated models are evaluated, but result in very modest improvements in accuracy. Thus, they do not seem justified in view of their increased complexity. The constant model with a MSE of 0.14 is a definite improvement over the Bulletin 17B skew map which reported a MSE of 0.302. Much of this improvement occurs because the GLS analysis correctly

reflects both the difference between the sampling error in at-site skew coefficient estimators and the precision of the regional model.

The sensitivity analysis for the B-GLS Southeastern U.S. skew analysis demonstrates how diagnostic statistics, and particularly leverage, can help in model selection. In cases where the data set has insufficient information to resolve the importance of a proposed independent variable in the regression, very large leverage statistics are observed. In particular, in the 342 site analysis, a proposed physiographic province whose affect is described by just three sites has leverage values between 30 and 45 times the threshold for large leverage. This example shows how leverage statistics can identify potentially unusual observations, which are overlooked when a data set is first organized.

The goal of the Southeastern U.S. study is to apply the improved regionalization methods to the estimation of a regional skewness coefficient, which is valuable in developing better estimates of flood frequency. Thus, the Southeastern U.S. data is used to compare flood frequency estimates using B-GLS regional skew to estimates using the *Bulletin 17B* regional skew. As is demonstrated in Chapter 4, the B-GLS regional skew estimation technique will allow for the regional skew to be weighted more heavily as it has a smaller MSE than the MSE provided by the regional skew from the Bulletin 17B skew map. Thus, the B-GLS regional skew method will lead to more accurate flood frequency estimation. The research presented in this thesis furthers the quasi-analytic Bayesian analysis of the Generalized Least Squares (GLS) regional hydrologic regression framework introduced by Reis *et al*. [2005] to more accurately and precisely estimate regional skewness coefficients. Examples are provided that illustrate both the performance of the B-GLS analysis in the estimation of regional skewness coefficients, and the value of the diagnostic statistics that have been developed.

5.3 Future Work

Future work will focus on developing improved cross-correlation models, as well as, enhanced screening procedures for redundant site pairs. The cross-correlation analysis presented in this thesis uses an OLS procedure. However, this framework neglects the variation and cross-correlation of the variances of the different residuals. Thus, the cross-correlation analysis can be improved by implementing both WLS and GLS procedures.

Currently, the screening metrics used to determine redundant site pairs relies solely on the distance between basin centroids and the basin drainage area. However, by employing estimates of the main channel length and width, as well as the distance between basin centroids and the basin drainage area, an improved understanding of basin geometry can be developed. This improved geometry can then be used to more accurately screen for redundant site pairs. These new techniques are expected to be tested on regional skew estimation studies in both California and Iowa, as well as in a nationwide study.

REFERENCES

Gruber, A.M., D.S. Reis Jr., and J. R. Stedinger (2007), Models of regional skew based on Bayesian GLS regression, *World Environmental & Water Resources Conference-Restoring out Natural Habitat*, edited by K.C. Kabbes, Tampa, Florida May 15-18, Paper 40927-3285.

Interagency Advisory Committee on Water Data, (1982), Guidelines for Determining Flood Flow Frequency, Bulletin #17B, U.S. Department of the Interior, U.S. Geological Survey, Office of Water Data Coordination, Reston Virginia.

Reis, D. S., Jr., J. R. Stedinger, and E. S. Martins (2005), Bayesian generalized least squares regression with application to log Pearson type 3 regional skew estimation, *Water Resour. Res.*, *41*, W10419, doi:10.1029/2004WR003445.

Stedinger, J.R. and V.W. Griffis, (2008), Flood Frequency Analysis in the United States: Time to Update. (editorial) *J. of Hydrol. Engineering*, April, pp. 199-204.

Tasker, G. D., and J. R. Stedinger (1989), An Operational GLS Model for Hydrologic Regression, *Journal of Hydrology*, 111, pp. 361-375.